1 Overview

2 Factor Analysis as a Mathematical Model

3 Matrix Algebra

4 Principal Components

4.1 Background

Principal components analysis is one of the venerable methods from the catalog of multivariate statistics. The goal of a principal component analysis is to create new variables that summarize as much information contained in many variables as possible. This is useful because the method reduces the number of variables one has to deal with.

PCA is often confused with factor analysis. You may have seen a "principal component factor analysis." It does not help that some statistical packages produce PCA when one requests that factor analysis be performed. The problem is that PCA actually is not factor analysis, not even a variety. PCA is data reduction whereas factor analysis is a latent variable model.

What is the difference?

In a nutshell, PCA starts with a set of variables that are all observed and then creates new variables, the components, that also are observed. There is nothing latent or unobserved about this operation. In contrast, factor analysis investigates unmeasured latent variables. A latent variable is a characteristic of people on a trait such as intelligence or sincerity or gregariousness. We never have the actual latent variable scores. People differ in intelligence but we do not have the latent variable itself. Hence the name, latent variable.

Definition: A component is a weighted combination of several variables. Define the original variables as, $(x_1, x_2, ..., x_p)$, and weights, $(q_1, q_2, ..., q_p)$. The result is

$$\boxed{z = q_1x_1 + q_2x_2 + \cdots + q_px_p}$$

Because of the equality of this expression, both the result, z, and the weighted combination are the component score.

Examples:

- The simple difference between a posttest and a pretest is a kind of component. For x_1 (pretest) and x_2 (posttest), set $q_1 = -1$, $q_2 = 1$ so that the difference is

$$z = x_2 - x_1$$

- The total score on a 20 item quiz is also a component. For most exams all items are given equal weight, say $q_j = 1$:

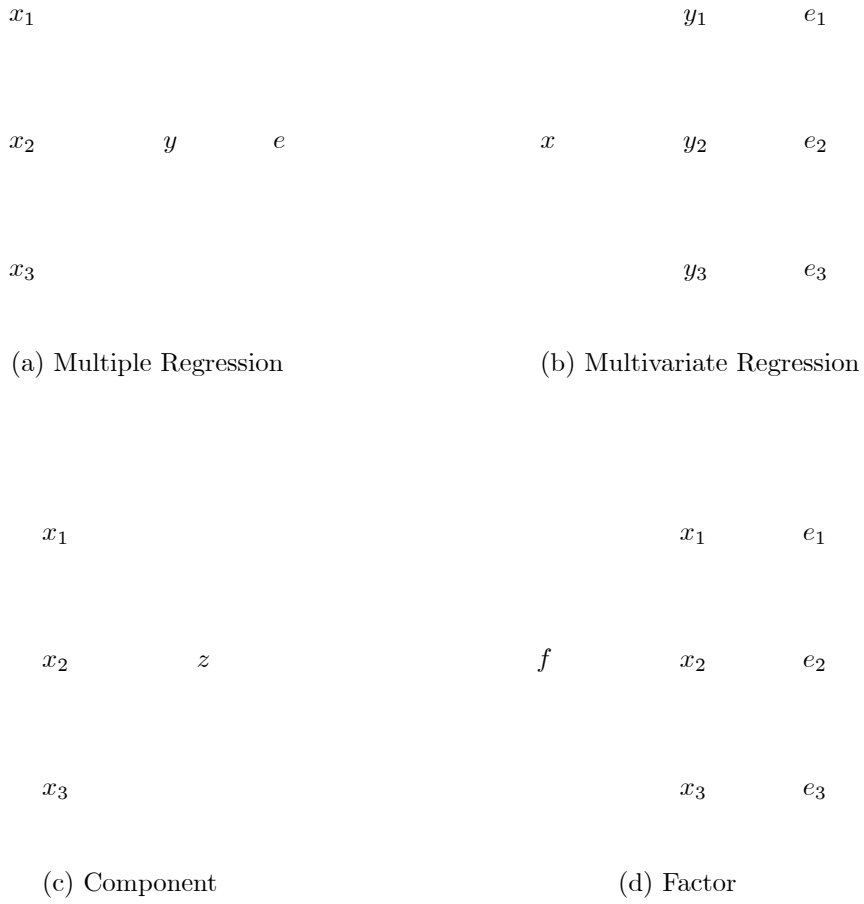
$$z = x_1 + x_2 + \cdots + x_{20}$$

- Many tests define the total score so that different parts receive different weight. For example, two true/false questions could receive 1/2 point each, two short answer questions might be weighted 1 point each, and two essay questions weighted 5 points each

$$z = \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 + x_3 + x_4 + 5x_5 + 5x_6$$

These examples all are components or weighted combinations of the x-variables.

What are the principal components? One way to explain is to contrast principal components with other procedures. Path diagrams are helpful. Here are four different procedures: two versions of regression, a component, and a factor. To keep it simple, assume that all variables are centered so that the means are zero.



(a) Multiple Regression (b) Multivariate Regression

(c) Component (d) Factor

Figure 1. Skeleton path diagrams for three procedures, (a) multiple regression, (b) multivariate regression, (c) a component, and (d) a genuine LV model with a single factor. The component z in panel (C) does not include any residual variable because z is completely determined by the x-variables.

- The first issue to note for the component is that all the variables are observed, including the new variable, z, the component. The latent variable, f, is not a measured variable that we have scores on. That's the major idea of a latent variable. It exists. People have scores on it. We don't have them
- The second idea is that the component is formed by differentially weighting each of the variables, x_1, x_2, and x_3. Thus z is itself an observed variable. The latent variable, f, exists apart from any manifest variable. So it makes sense that it serves as and independent variable in predicting x_j. That is a major conceptual distinction between a compenent and a latent variable.
- In constructing z from x_1, x_2, and x_3, there is no e in the diagram. The component is exactly equal to the weighted combination of the original variables. In the latent variable disgram each of the original variables is predicted from the factor, but some of the information in each x is unexplained. That is why the residual variables, e_1, e_2, e_3, are part of the diagram and also part of the factor analysis theory.

PCA is a nifty tool with many practical applications. It's too bad that components are often confused with factor analysis. In both there is the big summarizing variable, but that doesn't mean these are the same methods. We will investigate both methods so we can understand each on their own terms.

4.2 Example: Six Correlated Variables

Exercise physiologists conducted a study in which thirtythree normal volunteers were tested for their ability to move or pull their shoulder in a device that measures strength. There were three conditions defined by the starting angle of a lever the participants griped: 90%, 60%, and 30%. The measure of force applied to turning the device lever is torque. The data are in Table 1. One question was whether the angle of the lever was harder or easier to manipulate. Each score is the best of three trials attempted under each condition.

Correlations between the six measures of strength and age are in Table 2. The strength measures are highly correlated, ranging from 0.81 to 0.97. Young people are stronger than those who are older, which is show up as negative correlations between age and strength - younger age associated with greater strength.

Table 1. Measures of torque, the ability to move or pull a lever, using both the dominant and nondominant arm, based on resistance measured with a standard device. N = 33

			Ht	Wt	90^0		60^0		30^0	
		Age	(in.)	(lbs.)	DM	ND	DM	ND	DM	ND
Female	1	20	64	107	17	13	20	17	23	22
	2	23	68	140	25	25	28	29	31	31
	3	23	67	135	27	28	30	31	32	33
	4	23	67	155	23	28	27	29	27	32
	5	25	65	115	15	11	15	13	17	17
	6	26	68	147	27	17	25	21	32	27
	7	31	62	147	25	17	25	21	29	24
	8	31	66	137	19	15	17	17	21	19
	9	33	66	160	28	26	31	27	31	31
	10	36	66	118	23	23	26	27	27	25
	11	56	67	210	23	31	37	44	49	53
	12	59	67	130	15	17	17	19	20	20
	13	60	63	132	17	15	19	21	24	28
	14	60	64	180	15	15	17	19	19	21
	15	67	62	135	13	5	15	8	15	14
	16	73	62	124	11	9	13	13	19	17
Male	17	26	69	140	43	43	44	43	49	41
	18	28	71	175	45	43	48	45	53	52
	19	28	70	125	25	29	29	37	39	41
	20	28	70	175	39	41	49	47	55	44
	21	29	72	150	38	33	40	33	44	37
	22	30	68	145	53	41	51	40	59	44
	23	31	74	240	60	49	71	54	68	53
	24	32	67	168	32	31	37	31	39	30
	25	40	69	174	47	37	43	47	49	53
	26	41	72	190	33	25	29	25	39	27
	27	41	68	184	39	24	43	25	39	33
	28	56	70	200	21	11	23	12	33	24
	29	58	72	168	41	35	45	37	49	39
	30	59	73	170	31	32	31	31	35	38
	31	60	73	225	39	41	47	45	55	49
	32	68	67	140	31	23	33	27	37	33
	33	72	69	125	13	17	17	19	17	25

From Fisher & van Belle, 1993, Biostatistics, p. 452.

Table 2. Correlations between age and six measures of shoulder strength.

	Age	DM90	ND90	DM60	ND60	DM30	ND30
x: Age	1.000						
y_1: DM90	-0.339	1.000					
y_2: ND90	-0.332	0.895	1.000				
y_3: DM60	-0.287	0.958	0.926	1.000			
y_4: ND60	-0.292	0.837	0.967	0.899	1.000		
y_5: DM30	-0.254	0.931	0.919	0.964	0.915	1.000	
y_6: ND30	-0.194	0.812	0.912	0.871	0.960	0.910	1.000

The study investigated the association between age and strength. Because there are six response variables, we might examine six different regressions, using Age to predict each of the strength measures, y_j, $j = 1, ..., 6$, corresponding to DM90, ND90, ..., ND30. This is multivariate regression.

We usually write a model for data from a typical individual.

- Simple regression in scalar algebra is
$$y = \beta_0 + \beta_1 x + e$$
- With matrices, define the vector $\mathbf{x}$ (2 x 1) to be $\mathbf{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$ and the vector $\boldsymbol{\beta}$ (2 x 1) to be $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$. Simple regression in matrices is
$$y = \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} + e$$
$$\boldsymbol{\beta}' \qquad \mathbf{x} \qquad e$$
- For multivariate regression, there are m dependent variables, each predicted by one or more independent variables. This means that there are m sets of regression coefficients. Define the matrix of coefficients to be $\mathbf{B}$ of order (m x p), where p is the number of explanatory variables. For the example on p. 4 $\mathbf{B}$ is of order (6 x 2). The model is then
$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_6 \end{pmatrix} = \begin{pmatrix} \beta_{10} & \beta_{11} \\ \beta_{20} & \beta_{21} \\ \vdots & \vdots \\ \beta_{60} & \beta_{61} \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_6 \end{pmatrix}$$
$$\mathbf{y} \quad = \quad \mathbf{B} \quad \mathbf{x} \quad + \quad \mathbf{e}$$

For this particular problem predicting strength variables from age, a bare-bones path diagram for multivariate regression is below. Please draw in elements that should be included to make the figure complete.

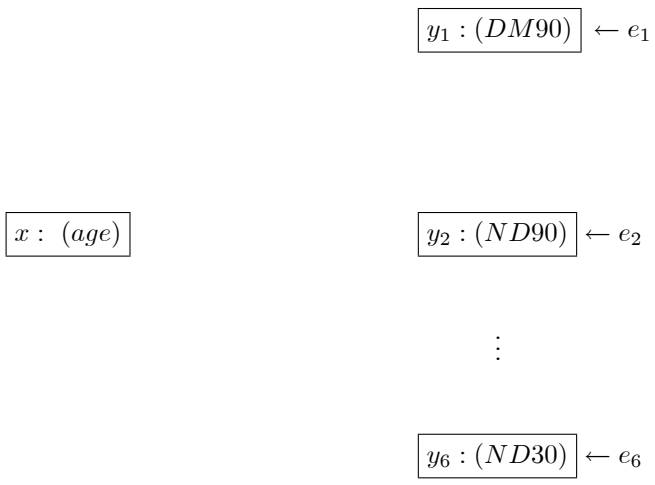


Table 3 shows estimated intercept and slope for each y_j when predicted from $x = Age$. The estimated slope for both DM90 and ND90 reach nominal significance ($\hat{\beta}_{j1} > |1.96 \cdot se(\hat{\beta}_{j1})|$), however the slopes for the other four measures of strength do not. Consequently, the conclusion about an effect when predicting Strength from Age is not straightforward. Age is significant for two measurements but not for four others.

Table 3. Estimated regression coefficients predicting each strength variable from $x =$ Age. Estimated slopes reach nominal significance for two of the six measures, y_1 and y_2 but for y_3 to y_6, slopes are not significant.

	y_1: DM90	y_2: ND90	y_3: DM60	y_4: ND60	y_5: DM30	y_6: ND30
$\hat{\beta}_{j0}$	39.2	35.1	41.0	37.4	44.3	38.1
$\hat{\beta}_{j1}(se)$	-0.25(.12)	-0.22(.11)	-0.23(.16)	-0.21(.12)	-0.21(.14)	-0.13(.12)

The problem of several correlated dependent variables can often be dealt with by summarizing. Instead of examining all the y variables in a multivariate regression, the six responses are summarized into one variable. This is what principal components does.

4.3 The Three Objectives of Principal Components

In regression the values of β_j are chosen to make the residual sum of squares, $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, where $\hat{y}_i = \beta x_i$, as small as possible. The coefficients or weights that accomplish the objectives of regression are denoted $\beta_1, \beta_p, ..., \beta_p$

Principal components also is based on a linear combination of the variables. Denote the sample of N individuals on p variables as $\mathbf{X}$ (N x p). Denote the PCA weights as $\mathbf{q} = (q_1, \ldots, q_p)$. For a single person, the linear combination of the x-variables in PCA is

$$\boxed{z_i = x_{i1}q_1 + \cdots + x_{ip}q_p} \tag{3.1}$$

For the entire sample of N individuals, the summary is

$$\mathbf{z} = \mathbf{Xq}$$

The principal components problem differs from regression because the objective of the linear combination is different. In particular there is no e in (3.1) because we create z from $x_1, ..., x_p$

Objective 1: (This is what we have been discussing). The PCA weights $\mathbf{q} = (q_1, \ldots, q_p)'$ are chosen so that z summarizes as much as possible the information contained in $\mathbf{x} = (x_1, \ldots, x_p)$'.

- The new variable, z, is called the *principal component of* $\mathbf{x}$
- The coefficients, $\mathbf{q}$, that produce the component scores are called the *principal component weights* or *loadings*.

Objective 2: The principal component score is constructed to simultaneously predict all p variables in $\mathbf{X}$ as well as possible . The weights to do this are $\mathbf{v} = (v_1, \ldots, v_p).'$ That is, z weighted by v_1 predicts x_1, z weighted by v_2 predicts x_2, and so on. The relationship is

$$\begin{aligned} \hat{\mathbf{X}}\ (N \text{ x } p) &= \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix} (v_1,\ v_2, \ldots, v_p) \\ &= (\mathbf{z}v_1,\ \mathbf{z}v_2, \ldots,\ \mathbf{z}v_p) \\ &= \mathbf{zv}' \end{aligned} \tag{3.2}$$

Each element of the predicted scores $\hat{\mathbf{X}}$ is as close as possible to the corresponding element of $\mathbf{X}$ so that the difference between actual and fitted scores

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}}$$

is small in the least squares sense over all p variables

$$\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2 = \sum_{i=1}^{N}\sum_{j=1}^{p}e_{ij}^2$$

where $\mathbf{E} = \{e_{ij}\}$.

It turns out that the weights $\mathbf{q}$ used to construct the original linear combination in (3.1) are the same as those needed for the multivariate regression problem in (3.2). So (i) summarizing the information in all variables of $\mathbf{X}$ into $\mathbf{z}$, and (ii) using the single variable $\mathbf{z}$ to predict all the variables in $\mathbf{X}$, are two parts of the same problem.

Objective 3: The principal component weights produce the variable z that has the largest possible sum of squares of any linear combination of $\mathbf{X}$. The sum of squares of z is defined

$$SS_z = \sum_{i=1}^{N} z_i^2 = \mathbf{z}'\mathbf{z} = \mathbf{q}'\mathbf{X}'\mathbf{X}\mathbf{q} \qquad \text{(a maximum)}$$

Because values of z can be made arbitrarily large just by making elements of $\mathbf{q}$ large, attention is restricted to weights that have sum of squares equal to 1

$$\sum q_j^2 = \mathbf{q}'\mathbf{q} = 1$$

Example Take a small data matrix, 5 x 2. Means of both columns are zero:

$$\mathbf{X} = \begin{pmatrix} 0 & -3 \\ 1 & 2 \\ -1 & -1 \\ 2 & 6 \\ -2 & -4 \end{pmatrix}$$

The problem is to find a weight vector with two elements, one for each column of $\mathbf{X}$, $\mathbf{q} = (q_1, q_2)'$, with $\mathbf{q}'\mathbf{q} = 1$, such that $\mathbf{z} = \mathbf{X}\mathbf{q}$ has maximal sum of squares, $SS_z = \mathbf{z}'\mathbf{z} = \mathbf{q}'\mathbf{X}'\mathbf{X}\mathbf{q}$. The sum of squares and cross-products matrix (SSCP) is

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 10 & 23 \\ 23 & 66 \end{pmatrix}$$

Here are four candidates for $\mathbf{q}$ with the sum of squares each produces:

Trial	$\mathbf{q}$		SS_z
1	0.7071	0.7071	61.0
2	0	1	66.0
3	$\sqrt{0.2}$	$\sqrt{0.8}$	73.1
4	0.337	0.941	74.2

The last set where $\mathbf{q} = (\mathbf{0}.337, 0.941)'$ is best for these data. It gives the largest sum of squares, $SS_z = 74.2$.

The three objects in this development have an interesting relationship. Define terms

$\mathbf{S} = \mathbf{X}'\mathbf{X}$	SSCP matrix computed from $\mathbf{X}$
$\mathbf{q}$	weights for constructing composite $\mathbf{z} = \mathbf{X}\mathbf{q}$
δ^2	sum of squares of $\mathbf{z}$ ($SS_z = \mathbf{z}'\mathbf{z} = \mathbf{q}'\mathbf{S}\mathbf{q}$)

The relationship is

(SSCP matrix times weights) equals (sum of squares times weights)

$$\begin{array}{cccccc} \mathbf{S} & \mathbf{q} & = & \delta^2 & \mathbf{q} & \\ \begin{pmatrix} 10 & 23 \\ 23 & 66 \end{pmatrix} & \begin{pmatrix} 0.34 \\ 0.94 \end{pmatrix} & = & 74.2 & \begin{pmatrix} 0.34 \\ 0.94 \end{pmatrix} & = \begin{pmatrix} 25.0 \\ 69.9 \end{pmatrix} \end{array}$$

In still other words, when $\mathbf{S}$ is multiplied by $\mathbf{q}$, $\mathbf{S}$ "behaves like" a scalar. The scalar that $\mathbf{S}$ behaves like is δ^2.

A vector and scalar, such as $\mathbf{q}$ and δ^2, that have this relationship to a symmetric matrix, $\mathbf{S}$

$$\boxed{\mathbf{S}\mathbf{q} = \mathbf{q}\delta^2} \tag{3.3}$$

are termed, respectively, an *eigenvector and eigenvalue of* $\mathbf{S}$. In general, if $\mathbf{S}$ is symmetric of order p, then $\mathbf{S}$ has p eigenvalues, denoted $(\delta_1^2, \ldots, \delta_p^2)$, and p sets of eigenvectors, $(\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_p)$, each $\mathbf{q}_j$ has p elements. The computational problem is to solve for the pairs $\{\mathbf{q}_j, \delta_j^2\}$, $j = 1, \ldots, p$, given an SSCP matrix such as our matrix $\mathbf{S}$.

The relation in (3.3) among $\mathbf{S}$, $\mathbf{q}$ and δ^2 is essentially the definition of eigenvalue/eigenvector pairs of $\mathbf{S}$. One use for them, described above, is to approximate the data matrix $\mathbf{X}$ by computing the scores of $\mathbf{z}$ which are then used to predict back to $\mathbf{X}$. There are *lots* of uses in other scientific fields.

To carry on with the example, we can use formula (3.1) to summarize as much information as possible from the matrix $\mathbf{X}$ into the vector $\mathbf{z}$

$$\begin{array}{ccccc} \mathbf{z} & = & \mathbf{X} & \mathbf{q} \\ \begin{pmatrix} -2.82 \\ 2.22 \\ -1.28 \\ 6.32 \\ -4.44 \end{pmatrix} & = & \begin{pmatrix} 0 & -3 \\ 1 & 2 \\ -1 & -1 \\ 2 & 6 \\ -2 & -4 \end{pmatrix} & \begin{pmatrix} .34 \\ .94 \end{pmatrix} \end{array}$$

Then with $\mathbf{z}$ available, one may predict back to the two variables in $\mathbf{X}$, using $\mathbf{z}$ as the predictor variable and $\mathbf{q}$ as weights

$$\begin{array}{ccccc} \hat{\mathbf{X}} & = & \mathbf{z} & \mathbf{q}' \\ \begin{pmatrix} -.95 & -2.66 \\ .75 & 2.09 \\ -.43 & -1.20 \\ 2.13 & 5.95 \\ -1.50 & -4.18 \end{pmatrix} & = & \begin{pmatrix} -2.82 \\ 2.22 \\ -1.28 \\ 6.32 \\ -4.44 \end{pmatrix} & \begin{pmatrix} .34 & .94 \end{pmatrix} \end{array}$$

The values in $\hat{\mathbf{X}}$ fit those in $\mathbf{X}$ better than any other matrix that results from a linear combination of a single independent variable. No other single independent variable can reproduce the two columns in $\mathbf{X}$ better than $\mathbf{z}$ can in conjunction with $\mathbf{q}$.

The residual matrix is

$$\begin{array}{ccccc} \mathbf{E} & = & \mathbf{X} & - & \hat{\mathbf{X}} \\ \begin{pmatrix} .95 & -.34 \\ .25 & -.09 \\ -.57 & .20 \\ -.13 & .05 \\ -.50 & .18 \end{pmatrix} & = & \begin{pmatrix} 0 & -3 \\ 1 & 2 \\ -1 & -1 \\ 2 & 6 \\ -2 & -4 \end{pmatrix} & - & \begin{pmatrix} -.95 & -2.66 \\ .75 & 2.09 \\ -.43 & -1.20 \\ 2.13 & 5.95 \\ -1.50 & -4.18 \end{pmatrix} \end{array}$$

The values in $\mathbf{E}$ are hopefully smallish. Check it out. This indicates that $\mathbf{X}$ is summarized to some degree by $\mathbf{z}$.

If the elements of $\mathbf{E}$ are not zero, then there is numerical information in $\mathbf{X}$ not accounted for by $\mathbf{z}$.

If $\mathbf{E}$ is not zero then we can define a second principal component that accounts for information in $\mathbf{X}$ not explained by the first principal component. Define $\mathbf{z}_1$ as the first component from above and $\mathbf{z}_2$ as the second component to be defined now. To calculate $\mathbf{z}_2$ we need the second eigenvalue/eigenvector pair from $\mathbf{S}$. Denote the first eigenvalue and eigenvector as δ_1^2, $\mathbf{q}_1$. The second pair is δ_2^2, $\mathbf{q}_2$. The second pair has the same relationship to $\mathbf{S}$ that the first pair did. In particular, the same equality holds that we saw earlier:

$$\begin{array}{cccccc} \mathbf{S} & \mathbf{q}_2 & = & \delta_2^2 & \mathbf{q}_2 & \\ \begin{pmatrix} 10 & 23 \\ 23 & 66 \end{pmatrix} & \begin{pmatrix} .94 \\ -.34 \end{pmatrix} & = & 1.77 & \begin{pmatrix} .94 \\ -.34 \end{pmatrix} & = \begin{pmatrix} 1.66 \\ -.60 \end{pmatrix} \end{array}$$

The second eigenvector applied to $\mathbf{X}$ produces the second principal component

$$\begin{array}{ccccc} \mathbf{z}_2 & = & \mathbf{X} & & \mathbf{q}_2 \\ \begin{pmatrix} 1.01 \\ .27 \\ -.60 \\ -.14 \\ -.54 \end{pmatrix} & = & \begin{pmatrix} 0 & -3 \\ 1 & 2 \\ -1 & -1 \\ 2 & 6 \\ -2 & -4 \end{pmatrix} & & \begin{pmatrix} .94 \\ -.34 \end{pmatrix} \end{array}$$

Because $\mathbf{X}$ has $p = 2$ columns there are only two eigenvalue/eigenvector pairs and therefore only two principal components. If there had been more than two variables in $\mathbf{X}$, we would carry on as above to calculate other components, $\mathbf{z}_3$, $\mathbf{z}_4$, ...

Both $\mathbf{z}_1$ and $\mathbf{z}_2$ summarize information in $\mathbf{X}$. Because they are artificial variables designed to explain $\mathbf{X}$, they are calculated so that information in $\mathbf{z}_2$ does not duplicate information already in $\mathbf{z}_1$. This requirement means that the two variates are uncorrelated.

$$corr(\mathbf{z}_1, \mathbf{z}_2) = 0$$

Substantively, this means the information in $\mathbf{z}_1$ is of a completely distinct kind than that in $\mathbf{z}_2$, at least in a statistical sense.

Now combine all these parts into a single package. Define the diagonal matrix of eigenvalues and the matrix of eigenvectors as

$$\mathbf{\Delta}^2 = diag(\delta_1^2, \delta_2^2) = \begin{pmatrix} 74.24 & \\ & 1.77 \end{pmatrix}$$

$$\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2) = \begin{pmatrix} .34 & .94 \\ .94 & -.34 \end{pmatrix}$$

The two principal components are

$$\begin{array}{ccc} \mathbf{Z} & = & \mathbf{X} \quad\quad \mathbf{Q} \\ \begin{pmatrix} -2.82 & 1.01 \\ 2.22 & .27 \\ -1.28 & -.60 \\ 6.32 & -.14 \\ -4.44 & -.54 \end{pmatrix} & = & \begin{pmatrix} 0 & -3 \\ 1 & 2 \\ -1 & -1 \\ 2 & 6 \\ -2 & -4 \end{pmatrix} \begin{pmatrix} .34 & .94 \\ .94 & -.34 \end{pmatrix} \end{array}$$

And these account for $\mathbf{X}$ completely. The fancy term is to say that $\mathbf{Z}$ and $\mathbf{Q}$ are a decomposition of $\mathbf{X}$ (Apt but not poetic - Defn: "decompose," synonym "rot").

$$\begin{array}{ccc} \hat{\mathbf{X}} & = & \mathbf{Z} \quad\quad \mathbf{Q}' \\ \begin{pmatrix} 0 & -3 \\ 1 & 2 \\ -1 & -1 \\ 2 & 6 \\ -2 & -4 \end{pmatrix} & = & \begin{pmatrix} -2.82 & 1.01 \\ 2.22 & .27 \\ -1.28 & -.60 \\ 6.32 & -.14 \\ -4.44 & -.54 \end{pmatrix} \begin{pmatrix} .34 & .94 \\ .94 & -.34 \end{pmatrix} \end{array}$$

The total sum of squares of $\mathbf{X}$ is a measure of how much information $\mathbf{X}$ contains. In the example,

$$\sum_{i=1}^{5}\sum_{j=1}^{2} x_{ij}^2 = 76$$

Of this amount, $\mathbf{z}_1$ accounts for $\sum_{i=1}^{5} z_{i1}^2 = 74.24$, while $\mathbf{z}_2$ only accounts for $\sum_{i=1}^{5} z_{i2}^2 = 1.77$. The two sums of squares equal the respective eigenvalues. That is the sum of squared elements in each $\mathbf{z}_j$ is broken down this way

$$\begin{aligned} \mathbf{Z}'\mathbf{Z} &= (\mathbf{XQ})'(\mathbf{XQ}) \\ &= \begin{pmatrix} 74.24 & \\ & 1.77 \end{pmatrix} \\ &= \mathbf{\Delta}^2 \end{aligned}$$

A practical implication of this (the main point for data analysis really) is that if we want to summarize the two scores in $\mathbf{X}$, the single principal component $\mathbf{z}_1$ contains most of the information of the original data. We get nearly the same information from $\mathbf{z}_1$ alone that was available in $\mathbf{x}_1$ and $\mathbf{x}_2$ together. Specifically, the proportion of variability of $\mathbf{X}$ explained by each component is

$$\delta_j^2 / trace(\mathbf{\Delta}^2)$$

In the example these are $74.28/76 = 0.977$ and $1.77/76 = 0.023$

The relationship between $\mathbf{S}$ and the set of eigenvalues and eigenvectors for the collection is

$$\begin{array}{ccccc} \mathbf{S} & \mathbf{Q} & = & \mathbf{Q} & \mathbf{\Delta}^2 \\ \begin{pmatrix} 10 & 23 \\ 23 & 66 \end{pmatrix} & \begin{pmatrix} .34 & .94 \\ .94 & -.34 \end{pmatrix} & = & \begin{pmatrix} .34 & .94 \\ .94 & -.34 \end{pmatrix} & \begin{pmatrix} 74.24 & \\ & 1.77 \end{pmatrix} \end{array}$$

Summary

1. The components of the matrix $\mathbf{X}$ (N x p) are the p linear combinations of $\mathbf{X}$

$$\begin{aligned}\mathbf{z}_1 &= \mathbf{X}\mathbf{q}_1\\ \mathbf{z}_2 &= \mathbf{X}\mathbf{q}_2\\ &\vdots\\ \mathbf{z}_p &= \mathbf{X}\mathbf{q}_p\end{aligned}$$

2. The principal components are the first few of these variates, say $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m$.

3. $\mathbf{z}_1$ has the largest sum of squares of any possible weighted combination of $\mathbf{X}$. Each subsequent component, $\mathbf{z}_k$, for $k = 2, \ldots, p$, has the largest sum of squares of any weighted combination of $\mathbf{X}$ which is uncorrelated with all $\mathbf{z}_k$, $k < j$.

4. The principal component weights are the eigenvectors, $\mathbf{q}_1, \ldots, \mathbf{q}_p$, of the SSCP matrix, $\mathbf{S} = \mathbf{X}'\mathbf{X}$, and the sum of squares of the jth component $\mathbf{z}_j$ is equal to the respective eigenvalue, $\delta_j^2 = \mathbf{z}_j'\mathbf{z}_j$.

5. The hallmark of an eigenvalue/eigenvector pair is that they form the so-called *characteristic equation* of $\mathbf{S}$

$$\mathbf{S}\mathbf{q}_j = \mathbf{q}_j\delta_j^2$$

These are the main formal ideas of principal components. Several other details to mention.

PCA of the 6 strength variables in Table 1
==========

```
options formdlim="_" nodate nocenter ls=160 ps=500;

data strength;

   input sex$ id age height weight dm1 nd1 dm2 nd2 dm3 nd3;
   if sex = 'Female' then g = 0;
                     else g = 1;

  datalines;
   Female   1   20  64  107  17  13  20  17  23  22
   Female   2   23  68  140  25  25  28  29  31  31
   Female   3   23  67  135  27  28  30  31  32  33
   Female   4   23  67  155  23  28  27  29  27  32
   Female   5   25  65  115  15  11  15  13  17  17
   Female   6   26  68  147  27  17  25  21  32  27
   Female   7   31  62  147  25  17  25  21  29  24
   Female   8   31  66  137  19  15  17  17  21  19
   Female   9   33  66  160  28  26  31  27  31  31
   Female  10   36  66  118  23  23  26  27  27  25
   Female  11   56  67  210  23  31  37  44  49  53
   Female  12   59  67  130  15  17  17  19  20  20
   Female  13   60  63  132  17  15  19  21  24  28
   Female  14   60  64  180  15  15  17  19  19  21
   Female  15   67  62  135  13   5  15   8  15  14
   Female  16   73  62  124  11   9  13  13  19  17
   Male    17   26  69  140  43  43  44  43  49  41
   Male    18   28  71  175  45  43  48  45  53  52
   Male    19   28  70  125  25  29  29  37  39  41
   Male    20   28  70  175  39  41  49  47  55  44
   Male    21   29  72  150  38  33  40  33  44  37
   Male    22   30  68  145  53  41  51  40  59  44
   Male    23   31  74  240  60  49  71  54  68  53
   Male    24   32  67  168  32  31  37  31  39  30
   Male    25   40  69  174  47  37  43  47  49  53
   Male    26   41  72  190  33  25  29  25  39  27
   Male    27   41  68  184  39  24  43  25  39  33
   Male    28   56  70  200  21  11  23  12  33  24
   Male    29   58  72  168  41  35  45  37  49  39
   Male    30   59  73  170  31  32  31  31  35  38
   Male    31   60  73  225  39  41  47  45  55  49
   Male    32   68  67  140  31  23  33  27  37  33
   Male    33   72  69  125  13  17  17  19  17  25
               ;

 * proc factor is the main factor analysis program in sas.  Lots of options.  It also does PCA

     o method=prin  produces the PCA
     o nfactor=1    requests one component
     o out=pca_res  produces a sas dataset containing the original variables plus the components ;

   proc factor data=strength method=prin n=1 out=pca_res(rename=(factor1=z));
     var dm1 -- nd3;
   run;
   proc print data=pca_res1;
   run;

* princomp also does "classical PCA" - no frills, few options;

   proc princomp data=strength N=1;
     var dm1 -- nd3;
   run;
```

```
The FACTOR Procedure
Initial Factor Method: Principal Components

Eigenvalues of the Correlation Matrix: Total = 6  Average = 1

        Eigenvalue    Difference    Proportion    Cumulative      < One component accounts for 92.7% of the
   1    5.55958486    5.29029877        0.9266        0.9266      < variability in all X-variables
   2    0.26928609    0.17896538        0.0449        0.9715
   3    0.09032071    0.04948702        0.0151        0.9865
   4    0.04083369    0.01434238        0.0068        0.9933
   5    0.02649131    0.01300798        0.0044        0.9978
   6    0.01348333                      0.0022        1.0000

1 factor will be retained by the NFACTOR criterion.


  Factor Pattern                                < The "pattern matrix" or "PCA loadings." In class notes this
                                                < is vector f, a column of the matrix F.  It is used to predict
            Factor1                             < matrix Xs from Matrix Zs:  Xs = Zs * f'
dm1         0.94040
nd1         0.97295
dm2         0.97297
nd2         0.96598
dm3         0.97652
nd3         0.94616


Variance Explained by Each Factor
   Factor1
 5.5595849

                     Final Communality Estimates: Total = 5.559585

       dm1             nd1             dm2             nd2             dm3             nd3
0.88435380      0.94662496      0.94667096      0.93312645      0.95358609      0.89522260

-------------------------------------------------------------------------------------------

The FACTOR Procedure
Initial Factor Method: Principal Components

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with Each Factor

   Factor1
 1.0000000


Standardized Scoring Coefficients               < Scoring coefficients.  In class notes, this is vector b, a
                                                < column of matrix B. It is used to create component scores
                                                < Zs from data Xs:  Zs = Xs * b
            Factor1

dm1         0.16915
nd1         0.17500
dm2         0.17501
nd2         0.17375
dm3         0.17565
nd3         0.17019
-----------------------------------------------------------------------------------------------
```

< Contents of dataset pca_res1 >

Obs	sex	id	age	height	weight	dm1	nd1	dm2	nd2	dm3	nd3	g	z
1	Female	1	20	64	107	17	13	20	17	23	22	0	-0.99994
2	Female	2	23	68	140	25	25	28	29	31	31	0	-0.19255
3	Female	3	23	67	135	27	28	30	31	32	33	0	-0.02166
4	Female	4	23	67	155	23	28	27	29	27	32	0	-0.22273
5	Female	5	25	65	115	15	11	15	13	17	17	0	-1.33183
6	Female	6	26	68	147	27	17	25	21	32	27	0	-0.49030
7	Female	7	31	62	147	25	17	25	21	29	24	0	-0.60028
8	Female	8	31	66	137	19	15	17	17	21	19	0	-1.05126
9	Female	9	33	66	160	28	26	31	27	31	31	0	-0.12628
10	Female	10	36	66	118	23	23	26	27	27	25	0	-0.44575
11	Female	11	56	67	210	23	31	37	44	49	53	0	0.76367
12	Female	12	59	67	130	15	17	17	19	20	20	0	-1.04389
13	Female	13	60	63	132	17	15	19	21	24	28	0	-0.82201
14	Female	14	60	64	180	15	15	17	19	19	21	0	-1.07219
15	Female	15	67	62	135	13	5	15	8	15	14	0	-1.59375
16	Female	16	73	62	124	11	9	13	13	19	17	0	-1.41784
17	Male	17	26	69	140	43	43	44	43	49	41	1	1.11806
18	Male	18	28	71	175	45	43	48	45	53	52	1	1.44136
19	Male	19	28	70	125	25	29	29	37	39	41	1	0.24853
20	Male	20	28	70	175	39	41	49	47	55	44	1	1.27706
21	Male	21	29	72	150	38	33	40	33	44	37	1	0.57608
22	Male	22	30	68	145	53	41	51	40	59	44	1	1.44289
23	Male	23	31	74	240	60	49	71	54	68	53	1	2.37396
24	Male	24	32	67	168	32	31	37	31	39	30	1	0.22766
25	Male	25	40	69	174	47	37	43	47	49	53	1	1.30514
26	Male	26	41	72	190	33	25	29	25	39	27	1	-0.08705
27	Male	27	41	68	184	39	24	43	25	39	33	1	0.25187
28	Male	28	56	70	200	21	11	23	12	33	24	1	-0.85331
29	Male	29	58	72	168	41	35	45	37	49	39	1	0.86425
30	Male	30	59	73	170	31	32	31	31	35	38	1	0.21966
31	Male	31	60	73	225	39	41	47	45	55	49	1	1.29636
32	Male	32	68	67	140	31	23	33	27	37	33	1	0.00063
33	Male	33	72	69	125	13	17	17	19	17	25	1	-1.03457

The principal component is z, a new variable that was created in this analysis. It summarizes the information contained in the six strength variables, dm1, nd1, dm2, ..., nd3. z does not contain all the information about the individual differences that the original six variables do. However it is the best summary of six-into-one that can be produced. In fact it picks up almost 93% of the information of six variables. A nice job of distillation.

How do you interpret z? The information z contains depends on how it is created. For this you examine the weights in what is labeled as the "factor pattern" in the results above. Later in these notes I'll denote the patterm matrix as the vector $\mathbf{f}$ (p x 1) (if there is just one component) or matrix $\mathbf{F}$ (p x m) (if there are $m \geq 2$ components). Rounded to two decimal points,

$$\mathbf{f} = \begin{pmatrix} .94 \\ .97 \\ .97 \\ .98 \\ .98 \\ .95 \end{pmatrix}$$

These weights show how strongly z is connected to x_j in the regression sense. If f_j is near zero, then z has almost nothing to do with x_j. If f_j is large, then z predicts x_j very well. In this example, all the weights are large and positive. This implies that z is very effective in explaining all the strength measures. We can essentially discard the original six variables and use only z without losing very much of the individual differences.

When each strength measure is predicted by z, we get really nice R^2 for all of them

```
proc reg data=pca_res ;
  model dm1 -- nd3 = z;
run;


Original
variable     R-square
  dm1         0.8844
  nd1         0.9466
  dm2         0.9467
  nd2         0.9331
  dm3         0.9536
  nd3         0.8952
```

Lets find the strongest and weakest woman. Its easy using the component because there is just one variable to examine. In general you use a component like any other variable: as a predictor in regression or as a dependent variable. If you want to know whether the mean scores on the component are significantly different for men and women, then a t-test using, for example, sex: $(F = 0, M = 1)$ and z.

For example, predict z from age. Recall in an earlier analysis that age could predict two of the six strength measures significantly, but the other four were N.S. - see Table 3. The new variable z summarizes all six strength measures, so perhaps age has a stronger relationship with z than with the individual variables.

For the regression $age \rightarrow z$: $\hat{R}^2 = 0.09$, $\hat{\beta}_1 = -0.017$. I was surprised by this.

```
proc reg data=pca_res;
  model z = age;
run;


The REG Procedure
Model: MODEL1
Dependent Variable: z

Number of Observations Read          33
Number of Observations Used          33

Root MSE              0.97108    R-Square     0.0865
Dependent Mean    -3.3643E-17    Adj R-Sq     0.0570
Coeff Var          -2.8864E18

                      Parameter Estimates
                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|
Intercept     1       0.72122        0.45367       1.59      0.1220
age           1      -0.01733        0.01012      -1.71      0.0967
-------------------------------------------------------------------------------------------
```

Better see a graph. Of course. Strength has a nonlinear relationship with age. Makes sense.

```
goptions reset=all hsize=8in vsize=7in;
symbol v=dot c=black i=none;
proc gplot data=pca_res;
plot z * age / vminor=0 hminor=0;
run; quit;
```

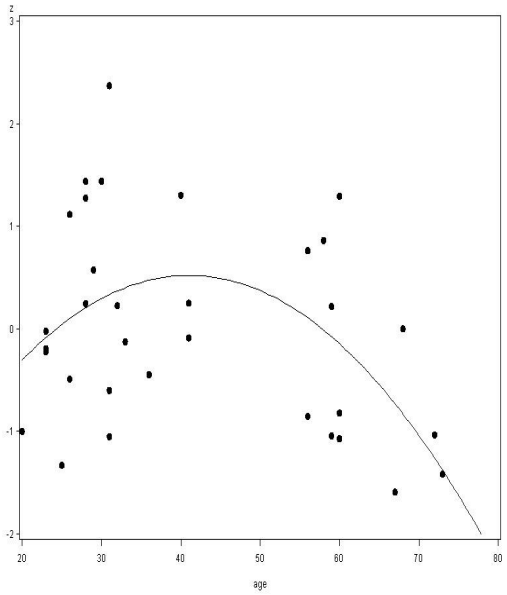


Figure 1. The first principal component, z_1, of six strength measures versus age. The correlation between z and age is $r_{z,age} = 0.294$. The correlation hides the fact that the two variables have a strong curvilinear relationship. The fitted line is the quadratic regression equation $z = \beta_0 + \beta_1 age + \beta_2 age^2 + e$

4.4 Multiple Components

One use of PCA is to construct a single component that summarizes the variables in a battery as well as one component can. Whether the component accounts for 30% of the variance or 85%, you use that single component. The component can subsequently be a predictor or response variable.

Much more often we want to account for an appreciable proportion of the variance of the battery. To do so two or more components are needed. When this is the goal, two new issues have to be addressed. The first, How many components should be used? The second, What is the meaning of the components in terms of the original variables?

Here is a new example with 10 variables and $N = 27$ participants. The variables are mood descriptors from a well regarded instrument called PANAS (Positive and Negative Affect Schedule). A person rates her- or himself on a scale of 1 to 5 (1 if you do not at all feel this mood today, 5 if this adjective is very much how you feel today). Five adjectives are positive and five are negative.

```
-------------------------------------------------------------------------------------------------------------

options formdlim='_' nocenter nodate ls=150 ps=500;
title 'N = 27 participants self-rated on the PANAS mood descriptors';

data mood;

  input id jittery  distress upset      afraid     scared
           inspired excited  determined interested enthusiastic;

datalines;
   1 2 1 1 1 2 2 1 2 2 2
   2 1 3 2 2 2 2 2 2 2 2
   3 1 1 1 2 1 1 1 2 2 1
   4 3 3 3 3 3 2 2 2 2 2
   5 2 1 1 2 2 4 4 5 4 4
   6 2 2 2 1 1 2 2 3 2 2
   7 1 2 1 2 2 2 1 3 3 3
   8 1 1 1 1 2 2 2 3 3 3
   9 3 3 3 3 3 1 2 1 2 1
  10 2 2 2 2 2 1 1 3 3 2
  11 1 1 1 1 1 3 3 4 4 3
  12 2 1 1 1 1 2 2 3 2 2
  13 3 2 4 3 2 1 2 2 1 1
  14 2 2 1 1 3 3 4 4 5 4
  15 1 2 3 2 1 2 4 3 1 4
  16 2 3 2 2 1 3 5 2 3 3
  17 3 4 2 2 2 1 2 2 1 2
  18 1 2 1 1 3 3 2 5 3 3
  19 4 4 4 2 3 3 1 1 2 1
  20 2 1 2 2 1 4 2 2 5 4
  21 2 1 1 2 1 3 4 3 2 4
  22 4 4 3 4 2 2 1 1 1 3
  23 2 3 3 2 2 3 4 3 3 2
  24 3 3 2 3 1 2 1 1 1 2
  25 2 1 2 1 1 2 3 1 2 2
  26 4 3 2 4 3 1 1 4 1 1
  27 1 1 1 1 2 3 2 1 3 3
     ;
```

```
+-------------------------------------------------------------------------+
|  * ODS graphics produce several different plots.                        |
|    A few other helpful options are selected also;                       |
|                                                                         |
|     ods html;                                                           |
|     ods graphics on;                                                    |
|        options ps=40 ls=70;                                             |
|        proc factor data=mood out=pca_res(rename=(factor1=z1 factor2=z2))|
|                    method=prin                                          |
|                    n=2                                                  |
|                    rotate=varimax                                       |
|                    plots=(scree(unpack) initloadings loadings)          |
|                    scree                                                |
|                    round                                                |
|                    flag=0.35                                            |
|                    reorder;                                             |
|        var jittery--enthusiastic;                                       |
|        run;                                                             |
|     ods graphics off;                                                   |
|     ods html close;                                                     |
|                                                                         |
|                                                                         |
|                                                                         |
|  * The PCA weights (the 'pattern') are equal to the correlations        |
|    between each component and the original variables - check it out;    |
|                                                                         |
|     options ls=150 ps=500;                                              |
|     proc corr data=pca_res;                                             |
|       var z1 z2;                                                        |
|       with jittery--enthusiastic;                                       |
|     run;                                                                |
|                                                                         |
|     proc print data=pca_res;                                            |
|     run;                                                                |
+-------------------------------------------------------------------------+
```

4.4.1 How Many Components?

This question has no satisfactory answer and in fact is not well posed (In mathematics, a problem is said to be well posed if a solution exists and if the solution is unique.) There are two sub-features that are always considered when deciding how many components to keep. Keep in mind that the decision about the best number is subjective.

The best anyone can do is to take the number of components that accounts for an "adequate" percentage of variance, and that make scientific sense, according to your informed judgment. It seems lame to emphasize subjective aspects of a quantitative and objective method. It would be nice if there were more respectable criteria. Sorry. People pretend this decision is objective and automatic. It's not.

Select Big Components Recall that the individual eigenvalues divided by the sum of eigenvalues is the proportion of variance that a component accounts for. In Table 3, components 1, 2 and 3 have eigenvalues of 4.61, 1.61, and 1.19. They individually account for 46%, 16%, and 12% of the total variance of the 10 variables. The cumulative percentage of variance explained by the components is 46%, 62% and 74%. So which are the "big" components and which are the small ones?

The idea is that the components that explain the most variance of a test battery are the important ones. Components that account for a small proportion are ignored.

The most popular rule for judging large components is easy. Not a bad way to start, although it lacks finesse and you should never take it as the final word. The rule is: The number of components to select are those that correspond to the number of eigenvalues great than 1.0. If you do not select the number of components some other way, most computer programs use the eigenvalue-greater-than-one rule. In the example, this rule leads to three components.

Table 3. Eigenvalues, proportion, and cumulative proportion of variance accounted for the complete set of principal components of the PANAS mood ratings.

	Eigenvalue	Difference	Proportion	Cumulative Proportion	
1	4.61349404	3.00832589	0.4613	0.4613	<- 1 component
2	1.60516815	0.41485525	0.1605	0.6219	<- 2 "
3	1.19031290	0.45134674	0.1190	0.7409	<- 3 "
4	0.73896616	0.16038148	0.0739	0.8148	
5	0.57858468	0.21107389	0.0579	0.8727	
6	0.36751078	0.07321086	0.0368	0.9094	
7	0.29429992	0.05090113	0.0294	0.9388	
8	0.24339880	0.01826492	0.0243	0.9632	
9	0.22513388	0.08200318	0.0225	0.9857	
10	0.14313070		0.0143	1.0000	

The Scree Plot and it's Elbow A famous graph is commonly used to show the magnitudes of the eigenvalues. It is called a scree plot, so-named after the boulders and stones of various sizes that accumulate at the bottom of a cliff. The plot displays the eigenvalues versus their order. Some people go through a solemn ritual while pondering this graph, acting all pious and scientific. As they study the figure they say they are conducting a

"scree test." That's pretentious. Nothing is being tested with this plot. The graph is fine and can help make this decision. But between you and me, it often soaks up more space in a report than it deserves.

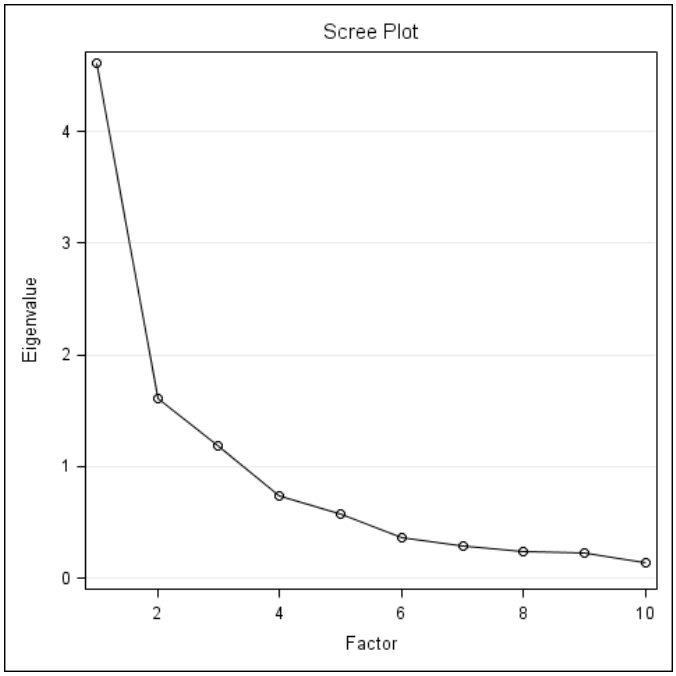


Figure 2. Graph of the magnitude of eigenvalues versus their order for the mood rating data, AKA the scree plot. This graph can help in deciding how many components are large, although it is not infallible.

Folk wisdom is to locate an elbow in the plot and take the number of large, important components to be where the bend occurs.

- A complicating fact with many sets of variables is that there is no distinct elbow, so you have to make a judgment about where the bend occurs. The scree plot often just shows that the decision is not at all clear-cut.
  - In this example, there is an elbow at 2 but another less prominent one at 4.
- Another complicating fact is that there is no guarantee, even if the large components are obvious, that the number of components above the elbow in the scree plot is in any way the best number. A satisfactory PCA solution is one that can be understood in light of the content of the variables. Deciding that one solution is understandable and others are murky is a matter of scientific judgment. Experienced people disagree.

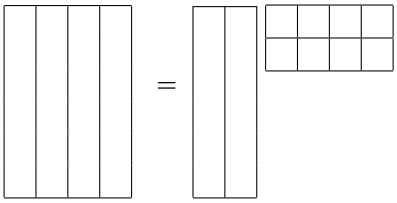
Consequently its a good idea to examine several solutions in differing numbers of components, starting at the elbow on the scree plot. It is almost always best to pick the solution that can be interpreted. This is more important than the amount of variance accounted for. I typically start examining solutions at the "elbow number of components" minus one. And then at the next largest number of components, and then also the next largest after that.

4.4.2 Interpretation: What is the Meaning of the Components?

PCA is especially helpful when the components have a clear connection to the original variables. Conversely, if the components make little sense in terms of the variables' content, then even a solution that accounts for practically all the variance may be unsatisfactory. Interpretation of results is a key aspect of PCA. It trumps the percentage of variance explained.

Interpretation is most straightforward when each component is associated with a few variables and unrelated to others. The key matrix is $\mathbf{Q}$, the (p x m) pattern or loadings matrix. The pattern matrix is used to reproduce the variables from the components.. This diagram has $p = 4$, $m = 2$.

$$\underset{N \text{ x } p}{\hat{\mathbf{X}}} = \underset{N \text{ x } m}{\mathbf{Z}} \quad \underset{m \text{ x } p}{\mathbf{Q}'} \tag{3.4}$$



We dearly want some elements in each column of $\mathbf{Q}$ to be large and others small.

The mood rating data has a double jointed scree plot in Figure 2. One elbow is at $m = 2$ components. There are positive and negative mood items so based on the content of the variables, a 2-component solution is reasonable. Let's start there.

Table 4 shows the pattern matrix with estimates listed two ways.

- On the left-hand side is the version as usually presented in research reports. It is difficult to process these estimates, first because the weights are presented with many more digits than a human being can assimilate, and second because there is no organization to the arrangement of the variables.
- To make the results easier to digest, it's a good idea to reorganize. The sas options to do this are 'round,' 'reorder' and 'flag = 0.35'
    1. Reflect all the weights in any column (that is multiply all entries by -1) so that positive weights indicate positive performance. It's also a good idea to reflect the weights so that the majority of values in any column are positive. I reflected column one so positive weights go with positive mood items.
    2. Drop the decimal point, round to two digits, and re-order variables in terms of their weights, listing from largest to smallest.
    3. Weights with $|w_{jk}| \leq 0.30$ or 0.35 are considered small. These variables can be skipped when re-ordering. Its actually difficult to reorder a table automatically. If your favorite computer program doesn't do it well enough, you may need to fine-tune by hand. The idea is to make the pattern matrix understandable for readers. Your eye can often see an arrangement of the table that tells the story best.
- A more effective version is on the right hand side of Table 4. Re-ordering the variables is good practice because the component affiliations identify important similarities and differences among the variables that were unknown before the analysis.

Don't underestimate the value of these stylistic conventions in presenting results.

Table 4. The pattern matrix presented two ways. On the left hand side, the weights are printed with many digits and in the original order of the variables. On the right the weights are rounded to two digits and variables are re-ordered in descending order of magnitude of weights. Re-ordering the variables puts those that are most important together on each component. The sas options to do this are 'round,' 'reorder' and 'flag.'

Pattern in almost unreadable form

	PC1	PC2
jittery	0.78377	0.38079
distress	0.75255	0.44923
upset	0.73809	0.35211
afraid	0.75204	0.30292
scared	0.36970	0.45108
inspired	-0.66499	0.55221
excited	-0.54256	0.47566
determined	-0.56062	0.23756
interested	-0.75550	0.32571
enthusiastic	-0.75113	0.38135

Pattern with rounded coefficients and re-ordered variables

	PC1	PC2
interested	76	33
enthusiastic	75	38
inspired	66	55
determined	56	24
excited	54	48
scared	-37	45
upset	-74	35
afraid	-75	30
distress	-75	45
jittery	-78	38

These are both weights and correlations.
Decimal points omitted.

The hope is that a few coefficients in each column are large while other coefficients for other variables are near zero. In this example, our dreams do not come true. A graph of q_1 versus q_2 is shown in Figure 3. Check out a few variables to translate from estimates of Table 4 to the plot of Figure 3.

- q_1 is a kind of contrast between positive and negative mood items. We might say the first component represents in a single variable both the healthy effects of positive traits and also the unhealthy effects of negative traits. If a person has endorsed mostly positive scores, they get a positive value on z_1 and if they endorse mostly negative items they get a negative z_1.
- This result is understandable. It is better if a component is associated with one group of variables and unrelated to others. Large loadings versus near zero loadings are easier to understand than are large positive versus large negative weights, because in the latter case we have to process additive contributions from some variables with subtractive contributions from others.
- q_2 has moderate loadings on all the items. I do not see an obvious interpretation for that.

This example is typical of many PCAs. Two components explain 62% of the variance of ten items, which is fine. But the interpretation of the components is strained and awkward. It a result that is satisfactory numerically but unclear in terms of scientific interpretation.

It often has been noted that PCA is designed to accomplish one job while the hope is that it accomplishes two. The only objective PCA understands is to explain as much of the information in the original variables as possible with a small number of components. It's a nifty trick in its own right. However the method knows nothing about the scientific clarity of the results – it just dutifully maximizes the variance that each component can explain.

Having said that, the configuration in Figure 4 is interesting because it shows clusters of the variables that are immediately recognizable: negative items all together, positive items all together, both groups clearly separated. The graph is an indication that a useful grouping of the variables exists. The problem is that raw principal components do not correspond to this clustering.

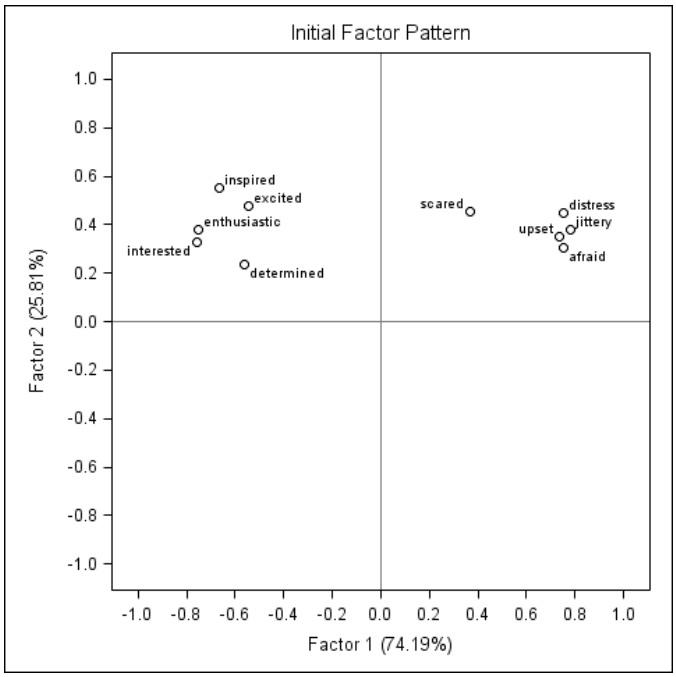


Figure 3. Plot of the pattern matrix for the principal components. The graph has the same information as the loadings of Table 4.

4.4.3 Nonprincipal Components

There is way out of the "PCA gives uninterpretable results" pickle. It can be called the method of nonprincipal components (NPCA). This is the method everyone uses when they report PCA results.

Recall for p variables that the $m < p$ largest principal components account for a proportion of the variability of the battery

$$\frac{\delta_1^2 + \delta_2^2 + \cdots + \delta_m^2}{p}$$

where δ_j^2 are the eigenvalues. For this problem based on Table 3, the first two principal components account for proportion

$$\frac{4.6135 + 1.6052}{10} = 0.622$$

Of these two the first, z_1, accounts for the proportion of $4.6135/10 = 0.4614$. This is the best that any single variable can achieve with these data. Similarly the second, z_2, accounts for the most information in $\mathbf{X}$ after the contribution of z_1 has been accounted for.

It turns out that there are many other sets of components that explain exactly the same proportion of variability in $\mathbf{X}$ that the principal components $\mathbf{Z}$ do. Denote these nonprincipal components as

$$\mathbf{Z}_r = (\mathbf{z}_{r_1},\ \mathbf{z}_{r_2}, \ldots, \mathbf{z}_{r_m})$$

$\mathbf{z}_{r_1}$ cannot account for as much of the variability of $\mathbf{X}$ as does $\mathbf{z}_1$. And $\mathbf{z}_{r_2}$ cannot account for the same variability in $\mathbf{X}$ that $\mathbf{z}_2$ does. However z_{r_1} and z_{r_2} together can explain exactly the same proportion of variance of $\mathbf{X}$ that z_1 and z_2 do. As $\mathbf{Q}$ is the set of weights associated with $\mathbf{Z}$, so there is associated with the nonprincipal components a weight matrix, $\mathbf{Q}_r$. The important relationship shown above in (3.4) is maintained with $\mathbf{Z}_r$ and $\mathbf{Q}_r$ exactly the same as with $\mathbf{Z}$ and $\mathbf{Q}$

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{Z}\mathbf{Q}' \\ &= \mathbf{Z}_r\mathbf{Q}_r' \end{aligned} \tag{3.5}$$

Why use NPCs when they perform the same as the PCs? Because they are specifically designed to produce interpretable components. And that can be invaluable.

A solution is cleanest if some weights are large and positive, and other weights close to zero. That happens when each of the component axes goes through a cluster of variables. In this example in Table 4 and Figure 3 there are two obvious clusters of variables. Neither axis hits a cluster. Consequently neither of the original components has a clear connection to these clusters of the variables.

4.4.4 Rotating Components

The search for an algorithm to find an informative $\mathbf{Q}_r$ with a simple pattern when the variables lie in distinct clusters is an interesting story. A key contribution was made in the 1950s by a clever grad student at Berkeley named Henry Kaiser.

Kaiser worked on this problem as a way to avoid studying for his comprehensive exams. A quirky eccentric, he also spray painted his sneakers before class with colors to reflect his mood, wrote scathing opinion pieces for the university newspaper on all kinds of subjects, was exuberantly Episcopalian, and ferociously conservative. In his research publications when he listed his university degrees, he always placed ES before MS or PhD. He was especially proud of being an Eagle Scout.

What did Kaiser do? He developed a computer algorithm that moves from a PCA solution that is messy and uninterpretable, like Figure 3, to a solution that is lovely and insightful, like the figure adjacent to Table 5 on the next page. Of course all this supposes that the correlation relationships among the variables really do make up cohesive clusters or understandable subgroups. If the data do not have a nice graphical structure, then that is the way the relationships are. We want the PCA to show us the nice groupings in the variables that exist.

Note that (3.5) is true as long as the configuration of variables in Figure 3 stays the same when the major axis lines are rotated. When we spin the axes, the relative positions of the variables in the figure stay the same. The coordinates of the variables, the magnitudes of the loadings, change as the axes change. Spinning changes the component scores. In this way (3.5) is true for the new $\mathbf{Q}_r$ and $\mathbf{Z}_r$ as it is for the original $\mathbf{Q}$ and $\mathbf{Z}$.

Spinning the axes of Figure 3 is called, naturally enough, a rotation of the configuration. Spinning the axes is represented as a matrix product that depends on the angle of rotation. To go from Figure 3 to the figure in Table 5 rotates the axes at an angle of about $\varphi = 60^o$ clockwise. Algebraically the two sets of changes are written

$$\mathbf{Q}_r = \mathbf{Q}\mathbf{T}(\varphi) \tag{3.6}$$

$$\mathbf{Z}_r = \mathbf{Z}\mathbf{T}(\varphi)'$$

where $\mathbf{T}(\varphi)$ is an (m x m) matrix that is calculated from φ

Kaiser's insight was figuring out that a good way to spin the axes is to calculate φ such that the loadings in $\mathbf{Q}_r$ have as large a variance as possible. To appreciate this, note that if all the loadings in $\mathbf{Q}_r$ are equal, then the variance of the loadings is zero. This is undesirable scientifically because it means that the component is connected to the variables to the same degree. On the other hand, as the loadings become different, the variance of the loadings becomes large. This is a desirable situation because then the components are differentially connected to the variables. Which is exactly what we want.

These ideas are the heart of the Kaiser's Varimax method, the most popular and successful way to uncover clusters in variables based on components. The function to maximize really is the variance of the squared loadings

$$V_{\max} = \frac{1}{p}\sum_{j=1}^{m}\sum_{i=1}^{p}\left([\mathbf{Q}_r]_{ij}^2 - \bar{q}_j\right)^2$$

where $\bar{q}_j$ is the mean of loadings in the jth column

$$\bar{q}_j = \frac{1}{p}\sum_{k=1}^{p}[\mathbf{Q}_r]_{kj}^2$$

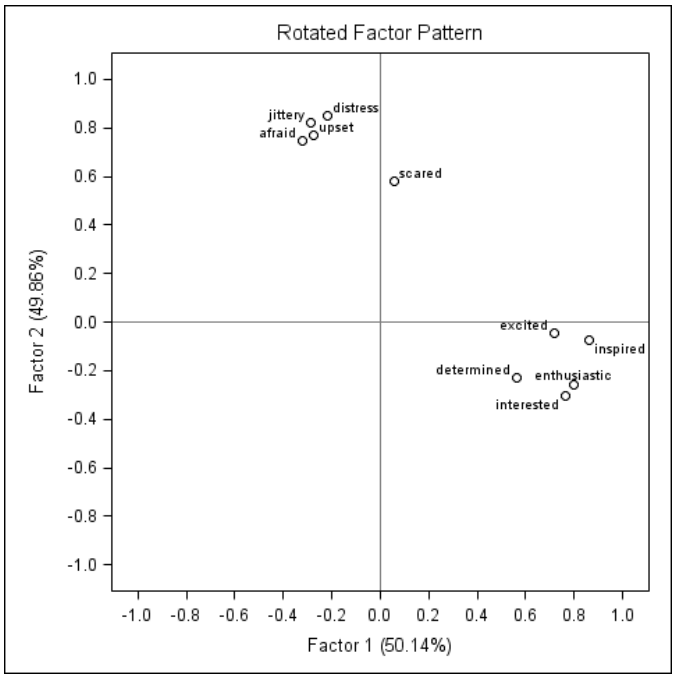
What is important to appreciate is that the rotated component solution is clearer than is the original principal component solution.

The rotated component solution for the mood variables is presented in Table 5 together with the corresponding plot. It is nicer than the initial solution shown in Figure 3. It is nicer because there are distinct clusters of variables and because an axis in the graph runs near each group.

Table 5. Rotated pattern matrix, $\mathbf{Q}_r$, with Varimax. The rotation works well here. The first component is associated with positive mood items, the second with negative items. The indication that the results are effective is that the variables fall into two distinct clusters and the axes run through both.

Pattern matrix with varimax rotation

	q_{r_1}	q_{r_2}
inspired	86	-08
enthusiastic	80	-26
interested	77	-30
excited	72	-05
determined	57	-23
distressed	-22	85
jittery	-29	82
upset	-28	77
afraid	-32	75
scared	-06	58



4.4.5 Check Other Solutions

I'm not unhappy with the two component solution for this problem. Having said that, its a good idea to examine other solutions for the variables to see whether a different number of components is more effective scientifically. Fewer components, in this case one, will explain less of the total variance, but if the solution seems clearer in terms of the content, then less variance explained may be tolerable. Of course three components will account for more of the total variance, but as a rule we don't necessarily want to explain more variance if the result is difficult to understand. Here is the old "use your judgment" issue again.

Table 6 has results for the one and three component solutions, the latter using Varimax rotation. The percentage of variance of the 10 items is 46% for one component, 74% for two. It would be great to explain more variance if it can be done clearly.

The one component solution is understandable. It is a single bi-polar scale. Individuals with a large positive score on z are more positive and less negative overall. Negative scores on the component occur because a person tends to endorse the negative rather than positive items.

The three component solution is hard to interpret. Yuck. In fact icky.

Given the content of the 10 items and the way that components in the three solutions perform, I prefer two.

Table 6. Estimated weights for the one- and three-component solutions for the 10 mood items. Varimax rotation for three components. Items have been re-ordered in terms of magnitude of the loadings and decimal points omitted. Weights that exceed |0.35| marked with an asterisk. Compared to the two component solution of Table 5, neither of these inspire me too much.

One component as SAS reports it		One component, weights reflected and sorted		Three rotated components		
jittery	78	interested	76	distress	85*	-15	13
distress	75	enthusiastic	75	upset	85*	-11	-18
afraid	75	inspired	66	jittery	84*	-22	10
upset	74	determined	56	afraid	78*	-24	2
scared	37	excited	54	inspired	-15	86*	11
excited	-54	scared	-37	excited	-5	82*	-15
determined	-56	upset	-74	enthusiastic	-31	81*	2
inspired	-66	afraid	-75	interested	-45*	62*	38*
enthusiastic	-75	distress	-75	scared	39*	-18	83*
interested	-76	jittery	-78	determined	-41*	33	59*

The goal of the enterprise is to understand how well the components account for variance and to understand how the components are related to the variables. In this problem there are no other variables to possibly predict or other statistical questions to investigate. The next page has all the data with the components attached. The first block sorts the cases in terms of their score on z_{r1}, and the second block shows the data with individuals sorted by z_{r2}. The scientific implication of this analysis is that there are two kinds of person, one who is cheerful and upbeat, and another who is negative and timid. Being cheerful does not mean a person has a negative outlook. The two types are not closely associated.

id	jittery	distress	upset	afraid	scared	inspired	excited	determined	interested	enthusias	Zr1	Zr2
5	2	1	1	2	2	4	4	5	4	4	2.20	0.07
14	2	2	1	1	3	3	4	4	5	4	2.12	0.27
20	2	1	2	2	1	4	2	2	5	4	1.31	-0.22
16	2	3	2	2	1	3	5	2	3	3	1.09	0.43
18	1	2	1	1	3	3	2	5	3	3	1.01	-0.33
23	2	3	3	2	2	3	4	3	3	2	0.95	0.90
11	1	1	1	1	1	3	3	4	4	3	0.88	-1.17
21	2	1	1	2	1	3	4	3	2	4	0.86	-0.51
15	1	2	3	2	1	2	4	3	1	4	0.36	-0.13
27	1	1	1	1	2	3	2	1	3	3	0.11	-1.02
8	1	1	1	1	2	2	2	3	3	3	0.05	-1.15
4	3	3	3	3	3	2	2	2	2	2	-0.06	1.39
7	1	2	1	2	2	2	1	3	3	3	-0.10	-0.68
19	4	4	4	2	3	3	1	1	2	1	-0.21	1.97
2	1	3	2	2	2	2	2	2	2	2	-0.40	-0.09
22	4	4	3	4	2	2	1	1	1	3	-0.45	1.84
6	2	2	2	1	1	2	2	3	2	2	-0.47	-0.68
12	2	1	1	1	1	2	2	3	2	2	-0.59	-1.26
25	2	1	2	1	1	2	3	1	2	2	-0.63	-0.91
10	2	2	2	2	2	1	1	3	3	2	-0.68	-0.32
1	2	1	1	1	2	2	1	2	2	2	-0.82	-1.05
17	3	4	2	2	2	1	2	2	1	2	-0.84	0.62
26	4	3	2	4	3	1	1	4	1	1	-0.85	1.37
9	3	3	3	3	3	1	2	1	2	1	-0.89	1.16
24	3	3	2	3	1	2	1	1	1	2	-1.10	0.30
13	3	2	4	3	2	1	2	2	1	1	-1.17	0.80
3	1	1	1	2	1	1	1	2	2	1	-1.69	-1.62

id	jittery	distress	upset	afraid	scared	inspired	excited	determined	interested	enthusias	Zr1	Zr2
19	4	4	4	2	3	3	1	1	2	1	-0.21	1.97
22	4	4	3	4	2	2	1	1	1	3	-0.45	1.84
4	3	3	3	3	3	2	2	2	2	2	-0.06	1.39
26	4	3	2	4	3	1	1	4	1	1	-0.85	1.37
9	3	3	3	3	3	1	2	1	2	1	-0.89	1.16
23	2	3	3	2	2	3	4	3	3	2	0.95	0.90
13	3	2	4	3	2	1	2	2	1	1	-1.17	0.80
17	3	4	2	2	2	1	2	2	1	2	-0.84	0.62
16	2	3	2	2	1	3	5	2	3	3	1.09	0.43
24	3	3	2	3	1	2	1	1	1	2	-1.10	0.30
14	2	2	1	1	3	3	4	4	5	4	2.12	0.27
5	2	1	1	2	2	4	4	5	4	4	2.20	0.07
2	1	3	2	2	2	2	2	2	2	2	-0.40	-0.09
15	1	2	3	2	1	2	4	3	1	4	0.36	-0.13
20	2	1	2	2	1	4	2	2	5	4	1.31	-0.22
10	2	2	2	2	2	1	1	3	3	2	-0.68	-0.32
18	1	2	1	1	3	3	2	5	3	3	1.01	-0.33
21	2	1	1	2	1	3	4	3	2	4	0.86	-0.51
7	1	2	1	2	2	2	1	3	3	3	-0.10	-0.68
6	2	2	2	1	1	2	2	3	2	2	-0.47	-0.68
25	2	1	2	1	1	2	3	1	2	2	-0.63	-0.91
27	1	1	1	1	2	3	2	1	3	3	0.11	-1.02
1	2	1	1	1	2	2	1	2	2	2	-0.82	-1.05
8	1	1	1	1	2	2	2	3	3	3	0.05	-1.15
11	1	1	1	1	1	3	3	4	4	3	0.88	-1.17
12	2	1	1	1	1	2	2	3	2	2	-0.59	-1.26
3	1	1	1	2	1	1	1	2	2	1	-1.69	-1.62

4.5 Recap

Principal components is not factor analysis. It is a method for data reduction. It begins with the actual variables collected in a project and leads to new variables that summarize as much information in the original battery as possible. There are no latent variables. The hope is that a few components summarize a large proportion of the variability of the variables. There is no statistical test for deciding how many components to retain, so judgment is required. Almost all principal components are rotated to an equivalent set of components that hopefully are more interpretable than are the principal components. Varimax rotation is by far the most popular method for this last step.

References

Cliff, N. (1987). *Analyzing multivariate data* (Ch. 13-14). New York: Harcourt, Brace, Jovanovich.

About as gentile a review of principal components as is available. Cliff covers many more topics than I have attempted in these notes.

Dunteman, G H (1989) *Principal components analysis.* Newbury Park CA: Sage

Available online for student use at OSU Library. Good review and not technical

Jackson, J E (2003) A user's guide to principal components. New York: Wiley

Jolliffe, I T (2008) *Principal component analysis.* New York: Springer

Complete treatment, technical

Morrison, D. F. (1990). *Multivariate statistical methods* (Ch. 8). New York: McGraw-Hill

Intermediate difficulty level, nice summary.

4.6 Eigenvalues and Eigenvectors

The principal components are based on standard results from the theory of eigenstructures. For a complete understanding of principal components, a more complete review of eigenstructures are summarized in this section.

1. If $\mathbf{S}$ (p x p) is a symmetric matrix, then the eigenstructure of $\mathbf{S}$ displays the relationship

$$\mathbf{SQ} = \mathbf{Q}\boldsymbol{\Delta}^2 \tag{4.4}$$

where $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_p]$ is a matrix of eigenvectors, each column of which is of length p, and the diagonal matrix $\boldsymbol{\Delta}^2 = diag(\delta_1^2, \ldots, \delta_p^2)$ contains eigenvalues.

Note on Rescaling: It is convenient to require that the eigenvalues in $\boldsymbol{\Delta}^2$ be arranged in descending order of magnitude, i.e., $(\delta_1^2 \geq \delta_2^2 \geq \delta_3^2 \cdots)$, and that the eigenvectors also be suitably rearranged. We further assume that the eigenvectors are *normalized*. That is, we assume that

$$\mathbf{q}_j'\mathbf{q}_j = 1 \qquad j=1,\ldots,p$$

If $\mathbf{Q}_0$ is a preliminary version of the eigenvector matrix which is not normalized, then $\text{Diag}(\mathbf{Q}_0'\mathbf{Q}_0) \neq \mathbf{I}$. The normalized eigenvectors can be easily obtained by first calculating $\mathbf{D} = \text{Diag}(\mathbf{Q}_0'\mathbf{Q}_0)$, and then rescaling $\mathbf{Q}_0$ by $\mathbf{Q} = \mathbf{Q}_0\mathbf{D}^{-\frac{1}{2}}$ to get the proper version of the eigenvector matrix that is normalized.

2. The eigenvectors of a symmetric matrix $\mathbf{S}$ (p x p) are *mutually orthogonal*; that is, $\mathbf{Q'Q} = \mathbf{I}$. Therefore the relationship in (4.4) is also written

$$\mathbf{S} = \mathbf{Q}\boldsymbol{\Delta}^2\mathbf{Q}'$$

in which $\mathbf{S}$ is decomposed into a triple product of two fundamental matrices, $\boldsymbol{\Delta}^2$ and $\mathbf{Q}$.

3. Problems in which eigenstructures play a part often begin with a matrix of observations, $\mathbf{X}$ (N x p), where $N > p$, and then use $\mathbf{X}$ to form other matrices. An important example of such a matrix is obtained by multiplying $\mathbf{X}$ by its transpose. This is called the *inner product* of $\mathbf{X}$. When $\mathbf{S}$ comes from $\mathbf{X}$ in this way, then the relationship is

$$\mathbf{S} = \mathbf{X'X} = \mathbf{Q}\boldsymbol{\Delta}^2\mathbf{Q}' \tag{4.5}$$

If $\mathbf{S}$ has this structure, all of its eigenvalues are positive or zero. If all the eigenvalues are positive, we say that $\mathbf{S}$ is a *positive definite* matrix. If one or more eigenvalues are zero, the matrix is called *positive semidefinite*.

4. The *rank* of a matrix, according to one of several equivalent definitions of matrix rank, is equal to the number of nonzero eigenvalues of the matrix. For any matrix $\mathbf{X}$ (N x p), the rank can never be greater than the smaller of N or p. When $\mathbf{X}$ is a data matrix, the rank is nearly always equal to p, the number of variables, unless a linear dependency exists among the columns, which would reduce the rank.

5. A second symmetric matrix can be constructed from $\mathbf{X}$, namely the *outer product* $\mathbf{XX'}$ which has order (N x N). This matrix also has an eigenstructure because it is symmetric. It is written

$$\mathbf{XX'} = \mathbf{B}\boldsymbol{\Lambda}^2\mathbf{B}' \tag{4.6}$$

where $\mathbf{B}$ consists of N columns $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_N]$, each of length N, and $\boldsymbol{\Lambda}^2 = diag(\lambda_1^2, \ldots, \lambda_N^2)$.

- The decomposition of $\mathbf{X'X}$ in (4.5) is related to the decomposition of $\mathbf{XX'}$ in (4.6). Partition $\mathbf{B}$ into two submatrices

$$\mathbf{B} = [\mathbf{P}(N \text{ x } p) \mid \mathbf{U}(N \text{ x } (N-p))]$$

where the first p columns are

$$\mathbf{P} = \mathbf{XQ}(\boldsymbol{\Delta}^2)^{-\frac{1}{2}} \tag{4.7}$$

and where $\mathbf{U}$ is arbitrary, generally taken as null.

- It is also the case that the eigenvalues from (4.6) in $\mathbf{\Lambda}^2$ have the form

$$\mathbf{\Lambda}^2 = \begin{pmatrix} \delta_1^2 & & & & & & \\ & \delta_2^2 & & & & & \\ & & \ddots & & & & \\ & & & \delta_p^2 & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix}$$

  That is, the non-zero eigenvalues of $\mathbf{XX'}$ are the same as the those of $\mathbf{X'X}$, and the first p eigenvectors of one matrix can be obtained from those of the other. The matrix $\mathbf{XX'}$ is N x N, with $N > p$. Thus there are N-p more eigenvalues and eigenvectors in $\mathbf{XX'}$ than in $\mathbf{X'X}$. Because of our definition of matrix rank, it must be that the last N-p eigenvalues of $\mathbf{XX'}$ are zero, and therefore the N - p last eigenvectors in $\mathbf{U}$ are arbitrary, generally taken as null.

- If $\mathbf{U}$ is a null matrix and if the last $N - p$ rows and columns of $\mathbf{\Lambda}^2$ are zero, then (4.6) can be written more compactly as

$$\mathbf{XX'} = \mathbf{P\Delta}^2\mathbf{P'}$$

Example:

Using $\mathbf{X}$ as defined earlier, the outer product and its decomposition are given below. The eigenvectors in $\mathbf{P}$ were obtained from (4.7) using $\mathbf{Q}$ and $\mathbf{\Delta}^2$ found earlier.

$$\underset{\mathbf{XX'}}{\begin{pmatrix} 9 & & & & \\ -6 & 5 & & & \\ 3 & -3 & 2 & & \\ -18 & 14 & -8 & 40 & \\ 12 & -10 & 6 & -28 & 20 \end{pmatrix}} = \underset{\mathbf{P}}{\begin{pmatrix} -.328 & .761 \\ .258 & .201 \\ -.148 & -.455 \\ .734 & -.105 \\ -.515 & -.402 \end{pmatrix}} \underset{\mathbf{\Delta}^2}{\begin{pmatrix} 74.24 & \\ & 1.77 \end{pmatrix}} \underset{\mathbf{P'}}{\begin{pmatrix} -.328 & .258 & -.148 & .734 & -.515 \\ .761 & .201 & -.455 & -.105 & -.402 \end{pmatrix}}$$

6. Every matrix $\mathbf{X}$ (N x p), in which N and p are completely arbitrary, can be decomposed into a triple product of fundamental matrices with the form,

$$\mathbf{X} = \mathbf{P\Delta Q'} \tag{4.9}$$

where the matrix $\mathbf{P}$ contains the first p eigenvectors of $\mathbf{XX'}$, and $\mathbf{Q}$ contains the eigenvectors of $\mathbf{X'X}$. This is known as the *singular value decomposition* or *spectral decomposition* of $\mathbf{X}$. The matrix $\mathbf{P}$ in this decomposition is sometimes called the matrix of *left-hand eigenvectors* of $\mathbf{X}$, $\mathbf{Q}$ is sometimes called the matrix of *right-hand eigenvectors*. The diagonal matrix, $\mathbf{\Delta} = (\delta_1, \ldots, \delta_p)$, contains the *singular values* of the matrix.

- The notational similarity between the eigenvalues, $\mathbf{\Delta}^2$, of $\mathbf{X'X}$ in (4.5) and the singular values, $\mathbf{\Delta}$, of $\mathbf{X}$ in (4.9) is not coincidental. The singular values of $\mathbf{X}$ and the eigenvalues of $\mathbf{X'X}$ are related by $\mathbf{\Delta}^2 = (\mathbf{\Delta})^2$.

Example:

The singular decomposition of $\mathbf{X}$ in our continuing example is

$$\underset{\mathbf{X}}{\begin{pmatrix} 0 & -3 \\ 1 & 2 \\ -1 & -1 \\ 2 & 6 \\ -2 & -4 \end{pmatrix}} = \underset{\mathbf{P}}{\begin{pmatrix} -.328 & .761 \\ .258 & .201 \\ -.148 & -.455 \\ .734 & -.105 \\ -.515 & -.402 \end{pmatrix}} \underset{\mathbf{\Delta}}{\begin{pmatrix} 8.616 & \\ & 1.328 \end{pmatrix}} \underset{\mathbf{Q'}}{\begin{pmatrix} .34 & .94 \\ .94 & -.34 \end{pmatrix}}$$

- This decomposition of $\mathbf{X}$ in (4.9) has a number of uses, not only for principal components analysis, but for many other multivariate methods as well. Most of the previous principles can be developed from it.
- For example, note that

$$\mathbf{X}'\mathbf{X} = (\mathbf{P}\mathbf{\Delta}\mathbf{Q}')'(\mathbf{P}\mathbf{\Delta}\mathbf{Q}') = \mathbf{Q}\mathbf{\Delta}\mathbf{P}'\mathbf{P}\mathbf{\Delta}\mathbf{Q}' = \mathbf{Q}\mathbf{\Delta}^2\mathbf{Q}'$$

which demonstrates that the eigenvalues of $\mathbf{X}'\mathbf{X}$ equal the squares of the singular values of $\mathbf{X}$, and that the eigenvectors of $\mathbf{X}'\mathbf{X}$ are the same as the right-hand eigenvectors of $\mathbf{X}$. A similar relationship holds for $\mathbf{X}\mathbf{X}'$, as can be seen by substitution.

The next three principles describe relationships that hold for decompositions in which only the first r largest singular values and the associated left- and right-hand eigenvectors of $\mathbf{X}$ are used. That is, although the complete sets of roots and vectors are available in $\mathbf{P}$ (N x p), $\mathbf{\Delta}$(p x p) and $\mathbf{Q}$ (p x p), we now use the subsets $\mathbf{P}_r$ (N x r), $\mathbf{\Delta}_r$ (r x r) and $\mathbf{Q}_r$ (p x r), where $r \leq p$.

7. The best approximation to the original matrix $\mathbf{X}$ is

$$\hat{\mathbf{X}}_r = \mathbf{P}_r\mathbf{\Delta}_r\mathbf{Q}_r' \tag{4.10}$$

in the sense that the sums of squares of the difference, $\mathbf{E} = \mathbf{X}$ - $\hat{\mathbf{X}}_r$, are small as possible. The matrix $\hat{\mathbf{X}}_r$ has dimensions n x p, but only has rank r. We say that $\hat{\mathbf{X}}_r$ "summarizes" all the information contained in $\mathbf{X}$ better than any other N x p matrix with rank r can. Another way to state this is to say that $\hat{\mathbf{X}}_r = \mathbf{P}_r\mathbf{\Delta}_r\mathbf{Q}_r'$ is the optimum *rank r matrix approximation* to $\mathbf{X}$.

Example:

Back to the example, the rank 1 approximation to $\mathbf{X}$ is

$$\begin{pmatrix} -.95 & -2.66 \\ .75 & 2.09 \\ -.43 & -1.20 \\ 2.13 & 5.95 \\ -1.50 & -4.18 \end{pmatrix} = \begin{pmatrix} -.328 \\ .258 \\ -.148 \\ .734 \\ -.515 \end{pmatrix} \quad 8.616 \quad \begin{pmatrix} .34 & .94 \end{pmatrix}$$

$$\hat{\mathbf{X}}_1 \qquad = \qquad \mathbf{P}_1 \qquad \mathbf{\Delta}_1 \qquad \mathbf{Q}_1'$$

8. The *principal components* of the matrix $\mathbf{X}$ are the columns of

$$\mathbf{Z}_r \ (N \text{ x } r) = \mathbf{X}\mathbf{Q}_r$$

where $r < p$ is generally chosen to be much smaller than p. A second form for $\mathbf{Z}_r$ is sometimes used. Write the matrices such that $\mathbf{P} = (\mathbf{P}_r, \mathbf{P}_k)$, $\mathbf{Q} = (\mathbf{Q}_r, \mathbf{Q}_k)$, and with $k = p - r$

$$\mathbf{\Delta} = \begin{pmatrix} \mathbf{\Delta}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}_k \end{pmatrix}$$

where $\mathbf{P}_r$, $\mathbf{Q}_r$, and $\mathbf{\Delta}_r$ are as defined above, while $\mathbf{P}_k$ is (N x (p-r)), $\mathbf{Q}_k$ is (p x (p-r)), and $\mathbf{\Delta}_k$ is [(p-r) x (p-r)]. Using the singular value decomposition of $\mathbf{X}$, and noting that $\mathbf{Q}_r'\mathbf{Q}_r = \mathbf{I}_r$ and that $\mathbf{Q}_k'\mathbf{Q}_r = \mathbf{0}$ (a k x r null matrix), we may also write the first r principal components as

$$\begin{aligned}\mathbf{Z}_r &= \mathbf{X}\mathbf{Q}_r = (\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}')\mathbf{Q}_r \\ &= (\mathbf{P}_r, \mathbf{P}_k)\begin{pmatrix} \boldsymbol{\Delta}_r & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta}_k \end{pmatrix}\begin{pmatrix} \mathbf{Q}_r' \\ \mathbf{Q}_k' \end{pmatrix}\mathbf{Q}_r \\ &= (\mathbf{P}_r, \mathbf{P}_k)\begin{pmatrix} \boldsymbol{\Delta}_r & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta}_k \end{pmatrix}\begin{pmatrix} \mathbf{I}_r \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{P}_r\boldsymbol{\Delta}_r \end{aligned} \tag{4.11}$$

This shows that the principal components may be computed directly from the primary matrices of this decomposition. Because of (4.11), another form for (4.10) is

$$\hat{\mathbf{X}}_r = \mathbf{Z}_r\mathbf{Q}_r' \tag{4.12}$$

Again, it can be seen that the matrix of reproduced scores in (4.10) or (4.12) has rank r because it is based on only r components in $\mathbf{Z}_r$ and $\mathbf{Q}_r$. The *residual matrix* containing that portion of X that is not reproduced by $\mathbf{Z}_r$, is

$$\mathbf{E}_r = \mathbf{X} - \hat{\mathbf{X}}_r$$

9. If $r = p$, then the residuals are zero: $\mathbf{E}_p = \mathbf{X} - \hat{\mathbf{X}}_p = \mathbf{0}$.

10. The principal components are uncorrelated, and the sums of squares of the jth principal component equals the jth eigenvalue of the matrix $\mathbf{X}'\mathbf{X}$. That is,

$$\mathbf{Z}_r'\mathbf{Z}_r = \boldsymbol{\Delta}_r^2$$

11. Because of the definition of $\mathbf{X}'\mathbf{X}$, its trace equals the total sum of squared elements in $\mathbf{X}$:

$$trace(\mathbf{X}'\mathbf{X}) = \sum_{j=1}^{p}\mathbf{x}_j'\mathbf{x}_j = \sum_{i=1}^{N}\sum_{j=1}^{p}x_{ij}^2$$

In addition,

(a) the total sum of squares of $\mathbf{X}$ equals the sum of the eigenvalues of $\mathbf{X}'\mathbf{X}$:

$$trace(\mathbf{X}'\mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{p}x_{ij}^2 = \sum_{k=1}^{p}\delta_k^2$$

(b) the total sum of squares of $\hat{\mathbf{X}}_r$ is $\sum_{j=1}^{r}\delta_j^2$

(c) the proportion of the total sum of squares of $\mathbf{X}$ accounted for by $\hat{\mathbf{X}}_r$ is the ratio of the sum of the first r eigenvalues to the sum of all eigenvalues

$$\frac{trace(\hat{\mathbf{X}}_r'\hat{\mathbf{X}}_r)}{trace(\mathbf{X}'\mathbf{X})} = \frac{\sum_{k=1}^{r}\delta_k^2}{\sum_{j=1}^{p}\delta_j^2}$$

4.7 Types of Variables

Multivariate methods are applied to different scalings of the original variables depending on the intent of the investigator or the characteristics of the method. Many writers use notation similar to the following for an N x p data matrix $\mathbf{X}$:

$\mathbf{X}$	Original data
$\mathbf{X}_d$	Centered (means zero)
$\mathbf{X}_s$	Standardized (means zero, variances unity)

The review of principal components to this point has concentrated on $\mathbf{X}_d$. Principal components can be defined for any of the three forms of the data; however there is no simple connection between any of the three sets of results. In practice, the principal components are generally computed from standardized variables. Because this is very common, we will follow suit and adapt the technology from above to standardized variables.

Another issue to clear up concerns the kind of output produced by popular statistical programs. Computer programs do not actually print the principal components when principal components are requested. The typical output is a weight matrix whose role must also be reviewed. After fighting through several pages of explanation about eigenvalues and eigenvectors, we will have to keep slugging through more material to get to the output your favorite program produces.

4.8 Standardized Principal Components from Standardized Variables

The singular value decomposition for $\mathbf{X}_s$ is

$$\mathbf{X}_s = \mathbf{P}\mathbf{\Delta}\mathbf{Q}'$$

Let $\mathbf{Q}_r$ contain the first r columns of $\mathbf{Q}$ with $r < p$. The first r principal components of $\mathbf{X}_s$ are

$$\mathbf{Z} = \mathbf{X}_s\mathbf{Q}_r$$

and the reproduced data matrix based on these scores is

$$\hat{\mathbf{X}}_s = \mathbf{Z}\mathbf{Q}_r'$$

These principal components are obtained from the standardized matrix $\mathbf{X}_s$. The columns of $\mathbf{Z}$ are uncorrelated, but are not standardized. In particular, the covariance matrix of $\mathbf{Z}$ is

$$\begin{aligned} cov(\mathbf{Z}) &= \frac{1}{N-1}\mathbf{Z}'\mathbf{Z} \\ &= \frac{1}{N-1}\mathbf{\Delta}_r^2 \\ &= \frac{1}{N-1}\begin{pmatrix} \delta_1^2 & & \\ & \ddots & \\ & & \delta_r^2 \end{pmatrix} \end{aligned}$$

This means that the variance of the k-th component, $\mathbf{z}_k$, equals the square of the k-th singular value divided by $N-1$.

Since the components are derived variables that exist for our use, we may as well have them in a form that is easy to interpret. Standardized components with variances of unity are nicer than unstandardized components with variances of $\delta_j^2/(N-1)$. Lets obtain standardized components. In virtually all applications, when one reads about "principal components," it is standardized components from standardized variables that are intended.

$$\mathbf{X}_s \rightarrow \mathbf{Z}_s$$

The diagonal matrix of standard deviations of the principal components is the square root of these elements

$$\mathbf{D}_z = \frac{1}{\sqrt{N-1}}\mathbf{\Delta}_r$$

To get standardized principal components, re-scale columns of $\mathbf{Z}$ by the appropriate standard deviations in $\mathbf{D}_z$. The standardized principal components will be denoted $\mathbf{Z}_s$

$$\begin{aligned} \mathbf{Z}_s &= \mathbf{Z}\mathbf{D}_z^{-1} \\ &= (\mathbf{X}_s\mathbf{Q}_r)\left(\tfrac{1}{\sqrt{N-1}}\mathbf{\Delta}_r\right)^{-1} \\ &= \mathbf{X}_s\mathbf{Q}_r\mathbf{\Delta}_r^{-1}\sqrt{N-1} \\ &= \mathbf{X}_s\mathbf{B} \end{aligned}$$

where $\mathbf{B} = \mathbf{Q}_r \mathbf{\Delta}_r^{-1} \sqrt{N-1}$ is a p x r matrix of weights for computing standardized principal components from the standardized variables. As opposed to the covariance matrix of the unstandardized components above, the covariance matrix of $\mathbf{Z}_s$ is

$$\begin{aligned} cov(\mathbf{Z}_s) &= \frac{1}{N-1} \mathbf{Z}_s' \mathbf{Z}_s \\ &= \mathbf{I}_r \end{aligned}$$

So $\mathbf{Z}_s$ is a little more appealing than $\mathbf{Z}$ because $\mathbf{Z}_s$ is uncorrelated *and also* has unit variances.

- What do computer programs typically print out? Not $\mathbf{Z}$, but not $\mathbf{Z}_s$ either (unless one specifically requests it as an output option). The weights $\mathbf{B}$ for producing $\mathbf{Z}_s$ from $\mathbf{X}_s$ might be interesting, but these are not the typical output. It is yet another weight matrix that is routinely produced.
- Earlier we made a big deal about the eigenvectors of $\mathbf{X}'\mathbf{X}$. The matrix $\mathbf{Q}$ plays two roles in the principal components of the centered data, $\mathbf{X}_d$. From item #8 above, $\mathbf{Q}$ produces $\mathbf{Z}$ from $\mathbf{X}_d$; but equally importantly, $\mathbf{Q}$ is used to approximate $\mathbf{X}_d$ from $\mathbf{Z}$

  $$\mathbf{Z} = \mathbf{X}_s \mathbf{Q}$$

  $$\hat{\mathbf{X}}_s = \mathbf{Z}\mathbf{Q}'$$

  When we want standardized components from standardized variables, $\mathbf{Q}$ does not give the components we need. Rather it is the matrix $\mathbf{B}$ above that does the first job: $\mathbf{B}$ produces standardized components from standardized variables. For the second job, a matrix we will denote as $\mathbf{F}$ contains the weights that approximate $\mathbf{X}_s$ from $\mathbf{Z}_s$. It is the matrix $\mathbf{F}$ that is generally the featured output from computer programs.
- So, to get the reproduced matrix $\mathbf{X}_s$ from $\mathbf{Z}_s$ we want this matrix $\mathbf{F}$. Note that standardized components can be expressed as

  $$\begin{aligned} \mathbf{Z}_s &= \mathbf{Z}\mathbf{D}_z^{-1} \\ &= \mathbf{Z}\mathbf{\Delta}_r^{-1}\sqrt{N-1} \end{aligned}$$

  This means that the unstandardized components can be obtained by re-arranging from above

  $$\mathbf{Z} = \mathbf{Z}_s \mathbf{\Delta}_r \tfrac{1}{\sqrt{N-1}} \tag{4.13}$$

  Next, recall that $\hat{\mathbf{X}}_s = \mathbf{Z}\mathbf{Q}_r'$. Substitute (4.13) for $\mathbf{Z}$ to give

  $$\begin{aligned} \hat{\mathbf{X}}_s &= \mathbf{Z}\mathbf{Q}_r' \\ &= \left(\mathbf{Z}_s \mathbf{\Delta}_r \tfrac{1}{\sqrt{N-1}}\right) \mathbf{Q}_r' \\ &= \mathbf{Z}_s \mathbf{F}' \end{aligned}$$

  And our hero is

  $$\boxed{\mathbf{F} = \mathbf{Q}_r \mathbf{\Delta}_r \tfrac{1}{\sqrt{N-1}}} \tag{4.14}$$

  The matrix $\mathbf{F}$ is of order p x k. It contains weights for reproducing values on the original standardized variables from the standardized principal components. It is often called the matrix of principal component "loadings." In many analyses, $\mathbf{F}$ is all that is reported about the components.

The following table gives the relationships between the standardized data $\mathbf{X}_s$ and two versions of the principal components. The first based on $\mathbf{Z}$ is simple, but the second pertaining to $\mathbf{Z}_s$ is by far the more common

alternative.

Summary of Matrices for Principal Components

$\mathbf{Z}$: Unstandardized	$\mathbf{Z}_s$: Standardized
$\mathbf{Z} = \mathbf{X}_s\mathbf{Q}_r$	$\mathbf{Z}_s = \mathbf{X}_s\mathbf{Q}_r\mathbf{\Delta}_r^{-1}\sqrt{N-1} = \mathbf{X}_s\mathbf{B}$
$\hat{\mathbf{X}}_s = \mathbf{Z}\mathbf{Q}_r'$	$\hat{\mathbf{X}}_s = \mathbf{Z}_s\mathbf{\Delta}_r\mathbf{Q}_r'\frac{1}{\sqrt{N-1}} = \mathbf{Z}_s\mathbf{F}'$
$Cov(\mathbf{Z}) = \mathbf{\Delta}_r^2\frac{1}{N-1}$	$Cov(\mathbf{Z}_s) = \mathbf{I}_r$

4.9 Principal Components from Correlations

Principal components are defined in term of the singular value decomposition of $\mathbf{X}_s$. However, most analyses in practice focus mainly, if not entirely, on the weight matrix $\mathbf{F}$ in (4.14). Interestingly if $\mathbf{F}$ is all that is needed, there is no need to compute the actual component scores, $\mathbf{Z}_s$. $\mathbf{F}$ is easily obtained from the sample correlation matrix, $\mathbf{R}$. The following steps show how this is derived (check the logic please)

$$\begin{aligned}\mathbf{R} &= \tfrac{1}{N-1}\mathbf{X}_s'\mathbf{X}_s \\ &= \tfrac{1}{N-1}(\mathbf{P\Delta Q}')'(\mathbf{P\Delta Q}') \\ &= \tfrac{1}{N-1}\mathbf{Q\Delta}^2\mathbf{Q}' \\ &= (\mathbf{Q\Delta}\tfrac{1}{\sqrt{N-1}})(\mathbf{Q\Delta}\tfrac{1}{\sqrt{N-1}})' \\ &= \mathbf{FF}'\end{aligned}$$

Of course, if the analysis is based only on $\mathbf{R}$, then the actual principal component scores, $\mathbf{Z}$ or $\mathbf{Z}_s$, are unavailable.

Summary of Decompositions

Singular value decomposition of $\mathbf{X}_s$	
$\mathbf{X}_s = \mathbf{P\Delta Q}'$	$\mathbf{\Delta}$: singular values of $\mathbf{X}_s$
Eigenstructure of $\mathbf{X}_s'\mathbf{X}_s$	
$\mathbf{X}_s'\mathbf{X}_s = \mathbf{Q\Delta}^2\mathbf{Q}'$	$\mathbf{\Delta}^2$: eigenvalues of $\mathbf{X}_s'\mathbf{X}_s$
Eigenstructure of $\mathbf{R}$	
$\mathbf{R} = \mathbf{QDQ}'$	$\mathbf{D} = \mathbf{\Delta}^2\frac{1}{N-1}$: eigenvalues of $\mathbf{R}$

4.10 Example

Determining whether $\mathbf{z}_1$ summarizes a "significant" proportion of the total variability of $\mathbf{X}_s$ is a matter of scientific judgement. In some problems a component analysis may be satisfactory if $\mathbf{z}_1$ reproduces 30% of the variability in $\mathbf{X}_s$. In others a researcher may be appalled if the variance accounted for by one component is only 80%. What decides the matter is the goal and purpose of the investigator.

The success of this summary cannot be decided on statistical grounds

The eigenvalues and eigenvectors were computed from the sample correlation matrix, using the identity $\mathbf{R} = \mathbf{QDQ}'$

Q Eigenvectors					**D** Eigenvalues				
.326	-.800	-.500	.055	-.031	3.42	0	0	0	0
.443	-.283	.687	-.501	.006	0	.833	0	0	0
.479	.055	.309	.816	.081	0	0	.413	0	0
.481	.367	-.343	-.244	.676	0	0	0	.261	0
.487	.377	-.256	-.142	-.732	0	0	0	0	.073

- The total variability in the correlation matrix among the five tests is equal to the sum of the eigenvalues. This value is $trace(\mathbf{D}) = 5$. The first principal component accounts for $100\left(\frac{3.42}{5}\right) = 68\%$ of the total variance, while two components account for $100\left(\frac{3.42+.833}{5}\right) = 85\%$. In this small example, one component accounting for two-thirds of the variability seems fine, in my humble opinion.

- The first *unstandardized* principal component is $\mathbf{z}_1 = \mathbf{X}_s\mathbf{q}_1$, where

$$\mathbf{q}_1 = (.326, .443, .479, .481, .487)'$$

  is the first column of the matrix $\mathbf{Q}$ above.

- The first *standardized* principal component is $\mathbf{z}_{s1} = \mathbf{X}_s\left(\mathbf{q}_1/\sqrt{d_{11}}\right)$. The weights to be applied to $\mathbf{X}_s$ to obtain the first standardized principal component are

$$\frac{\mathbf{q}_1}{\sqrt{d_{11}}} = \begin{pmatrix} .326 \\ .443 \\ .479 \\ .481 \\ .487 \end{pmatrix} \frac{1}{\sqrt{3.42}} = \begin{pmatrix} .176 \\ .240 \\ .259 \\ .260 \\ .263 \end{pmatrix}$$

  Values of $\mathbf{z}_{s1}$ for the ten cases selected from the larger sample are listed in the last column of the table that contains the sample data.

- To approximate the five original variables in $\mathbf{X}_s$ using only the first standardized principal component, $\mathbf{z}_{s1}$, that is, to obtain the rank one approximation $\hat{\mathbf{X}}_s$ to $\mathbf{X}_s$, we use the "loadings" for the first principal component. The complete set of loadings is $\mathbf{F} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}$

$\mathbf{F}$, Principal Component Loadings				
.603	-.730	-.322	.028	-.088
.820	-.259	.442	-.256	.002
.885	.050	.199	.417	.022
.889	.335	-.221	-.125	.183
.900	.344	-.165	-.072	-.198

- The reproduced matrix using only the first component is $\hat{\mathbf{X}}_s = \mathbf{z}_{s1}\mathbf{f}_1'$, where $\mathbf{f}_1$ is the first column of $\mathbf{F}$.

- $\mathbf{f}_1$ contains weights to apply to $\mathbf{z}_{s1}$ to reproduce $\mathbf{X}_s$. Researchers attempt to "interpret" the magnitudes to these weights to understand how the component is related to the original variables. Examining $\mathbf{f}_1$ it can be seen that $\mathbf{z}_{s1}$ is relatively effective in predicting all five of the variables because the coefficients are moderate to large in size. The value for X_{s5} is $f_{51} = .900$, meaning that $\mathbf{z}_{s1}$ is strongly associated with X_{s5}.

Example

In a famous study, Skodak and Skeels (1949, *J. Genetic Psych., 75*, 85-125) presented data from a longitudinal study of $N = 100$ adopted children who were contacted on several occasions from age 3 to 16. Five intelligence tests were administered to this sample. A small collection of the scores for five of the boys and five girls are in Table 1. Means, standard deviations and correlations are also listed.

Table 1. Five IQ measures for 10 children with the principal component that best summarizes the five test scores

	Intelligence Tests					
	1	2	3	4	5	$\mathbf{z}_{s1}$
1-B	121	114	115	105	115	.047
2-B	120	115	109	106	117	-.063
3-B	131	109	113	95	106	-.556
4-B	102	121	139	114	132	1.50
5-B	126	115	113	90	99	-.870
1-G	90	94	95	77	90	-3.79
2-G	104	114	104	100	108	-1.15
3-G	125	124	116	127	134	1.83
4-G	99	102	128	126	142	1.13
5-G	135	112	118	118	124	1.15

Correlations					
	1				
	.534	1			
	.444	.694	1		
	.357	.577	.712	1	
	.344	.595	.747	.924	1
Mean	116.5	112.3	114.9	107.1	116.8
S.D.	13.2	13.7	13.5	14.6	15.4

The five IQ scores are correlated to varying degrees, and are certainly not independent. Rather than maintain all of them throughout a complicated series of analyses, it may be preferable to acknowledge that there is considerable redundancy in the collection, and attempt to summarize as much information as possible by one principal component. If this can be done effectively, then $\mathbf{z}_1$ may replace $x_1, \ldots, x_5$ in subsequent analyses.