Version: August 4, 2011

Notes on Principal Component Analysis

A preliminary introduction to principal components was given in our brief discussion of the spectral decomposition (i.e., the eigenvector/eigenvalue decomposition) of a matrix and what it might be used for. We will now be a bit more systematic, and begin by making three introductory comments:

(a) Principal component analysis (PCA) deals with only one set of variables without the need for categorizing the variables as being independent or dependent. There is asymmetry in the discussion of the general linear model; in PCA, however, we analyze the relationships among the variables in one set and *not* between two.

(b) As always, everything can be done computationally without the Multivariate Normal (MVN) assumption; we are just getting descriptive statistics. When significance tests and the like are desired, the MVN assumption becomes indispensable. Also, MVN gives some very nice interpretations for what the principal components are in terms of our constant density ellipsoids.

(c) Finally, it is probably best if you are doing a PCA, not to refer to these as "factors". A lot of blood and ill-will has been spilt and spread over the distinction between component analysis (which involves linear combinations of *observable* variables), and the estimation of a factor model (which involves the use of underlying latent variables or factors, and the estimation of the factor structure). We will get sloppy ourselves later, but some people really get exercised about these things.

We will begin working with the population (but everything translates more-or-less directly for a sample):

Suppose $[X_1, X_2, \ldots, X_p] = \mathbf{X}'$ is a set of p random variables, with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{\Sigma}$. I want to define p linear combinations of $\mathbf{X}'$ that represent the information in $\mathbf{X}'$ more parsimoniously. Specifically, find $\mathbf{a}_1, \ldots, \mathbf{a}_p$ such that $\mathbf{a}_1'\mathbf{X}, \ldots, \mathbf{a}_p'\mathbf{X}$ gives the same information as $\mathbf{X}'$, but the new random variables, $\mathbf{a}_1'\mathbf{X}, \ldots, \mathbf{a}_p'\mathbf{X}$, are "nicer".

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the p roots (eigenvalues) of the matrix $\mathbf{\Sigma}$, and let $\mathbf{a}_1, \ldots, \mathbf{a}_p$ be the corresponding eigenvectors. If some roots are not distinct, I can still pick corresponding eigenvectors to be orthogonal. Choose an eigenvector $\mathbf{a}_i$ so $\mathbf{a}_i'\mathbf{a}_i = 1$, i.e., a normalized eigenvector. Then, $\mathbf{a}_i'\mathbf{X}$ is the i^{th} principal component of the random variables in $\mathbf{X}'$.

Properties:

1) $\text{Var}(\mathbf{a}_i'\mathbf{X}) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i = \lambda_i$

We know $\mathbf{\Sigma}\mathbf{a}_i = \lambda_i\mathbf{a}_i$, because $\mathbf{a}_i$ is the eigenvector for λ_i; thus, $\mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i = \mathbf{a}_i'\lambda_i\mathbf{a}_i = \lambda_i$. In words, the variance of the i^{th} principal component is λ_i, the root.

Also, for all vectors $\mathbf{b}_i$ such that $\mathbf{b}_i$ is orthogonal to $\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}$, and $\mathbf{b}_i'\mathbf{b}_i = 1$, $\text{Var}(\mathbf{b}_i'\mathbf{X})$ is the greatest it can be (i.e., λ_i) when $\mathbf{b}_i = \mathbf{a}_i$.

2) $\mathbf{a}_i$ and $\mathbf{a}_j$ are orthogonal, i.e., $\mathbf{a}_i'\mathbf{a}_j = 0$

3) $\mathrm{Cov}(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_j = \mathbf{a}_i'\lambda_j\mathbf{a}_j = \lambda_j\mathbf{a}_i'\mathbf{a}_j = 0$

4) $\mathrm{Tr}(\mathbf{\Sigma}) = \lambda_1 + \cdots + \lambda_p =$ sum of variances for all p principal components, and for $X_1, \ldots, X_p$. The importance of the i^{th} principal component is

$$\lambda_i/\mathrm{Tr}(\mathbf{\Sigma}) \ ,$$

which is equal to the variance of the i^{th} principal component divided by the total variance in the system of p random variables, $\mathbf{X}_1, \ldots, \mathbf{X}_p$; it is the proportion of the total variance explained by the i^{th} component.

If the first few principal components account for most of the variation, then we might interpret these components as "factors" underlying the whole set $\mathbf{X}_1, \ldots, \mathbf{X}_p$. This is the basis of *principal factor analysis*.

The question of how many components (or factors, or clusters, or dimensions) usually has no definitive answer. Some attempt has been made to do what are called Scree plots, and graphically see how many components to retain. A plot is constructed of the value of the eigenvalue on the y-axis and the number of the eigenvalue (e.g., 1, 2, 3, and so on) on the x-axis, and you look for an "elbow" to see where to stop. Scree or talus is the pile of rock debris (detritus) at the foot of a cliff, i.e., the sloping mass of debris at the bottom of the cliff. I, unfortunately, can never see an "elbow"!

If we let a population correlation matrix corresponding to $\mathbf{\Sigma}$ be denoted as $\mathbf{P}$, then $\mathrm{Tr}(\mathbf{P}) = p$, and we might use only those principal components that have variance of $\lambda_i \geq 1$ — otherwise, the component would "explain" less variance than would a single variable.

A major rub — if I do principal components on the correlation matrix, $\mathbf{P}$, *and* on the original variance-covariance matrix, $\mathbf{\Sigma}$, the structures obtained are generally different. This is one reason the "true believers" might prefer a factor analysis model over a PCA because the former holds out some hope for an invariance (to scaling). Generally, it seems more reasonable to always use the population correlation matrix, $\mathbf{P}$; the units of the original variables become irrelevant, and it is much easier to interpret the principal components through their coefficients.

The j^{th} principal component is $\mathbf{a}_j'\mathbf{X}$:

$\text{Cov}(\mathbf{a}_j'\mathbf{X}, X_i) = \text{Cov}(\mathbf{a}_j'\mathbf{X}, \mathbf{b}'\mathbf{X})$, where $\mathbf{b}' = [0 \cdots 0\ 1\ 0 \cdots 0]$, with the 1 in the i^{th} position, $= \mathbf{a}_j'\mathbf{\Sigma}\mathbf{b} = \mathbf{b}'\mathbf{\Sigma}\mathbf{a}_j = \mathbf{b}'\lambda_j\mathbf{a}_j = \lambda_j$ times the i^{th} component of $\mathbf{a}_j = \lambda_j a_{ij}$. Thus, $\text{Cor}(\mathbf{a}_j'\mathbf{X}, X_i) =$

$$\frac{\lambda_j a_{ij}}{\sqrt{\lambda_j}\sigma_i} = \frac{\sqrt{\lambda_j}a_{ij}}{\sigma_i} \ ,$$

where σ_i is the standard deviation of X_i. This correlation is called the *loading* of X_i on the j^{th} component. Generally, these correlations can be used to see the contribution of each variable to each of the principal components.

If the population covariance matrix, $\mathbf{\Sigma}$, is replaced by the sample covariance matrix, $\mathbf{S}$, we obtain sample principal components; if the population correlation matrix, $\mathbf{P}$, is replaced by the sample correlation matrix, $\mathbf{R}$, we again obtain sample principal components. These structures are generally different.

The covariance matrix $\mathbf{S}$ (or $\mathbf{\Sigma}$) can be represented by

$$\mathbf{S} = [\mathbf{a}_1, \ldots, \mathbf{a}_p] \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda_p} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda_p} \end{bmatrix} \begin{bmatrix} \mathbf{a}'_1 \\ \vdots \\ \mathbf{a}'_p \end{bmatrix} \equiv \mathbf{L}\mathbf{L}'$$

or as the sum of p, $p \times p$ matrices,

$$\mathbf{S} = \lambda_1 \mathbf{a}_1 \mathbf{a}'_1 + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}'_p \ .$$

Given the ordering of the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, the least-squares approximation to $\mathbf{S}$ of rank r is $\lambda_1 \mathbf{a}_1 \mathbf{a}'_1 + \cdots + \lambda_1 \mathbf{a}_r \mathbf{a}'_r$, and the residual matrix, $\mathbf{S} - \lambda_1 \mathbf{a}_1 \mathbf{a}'_1 - \cdots - \lambda_1 \mathbf{a}_r \mathbf{a}'_r$, is $\lambda_{r+1} \mathbf{a}_{r+1} \mathbf{a}'_{r+1} + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}'_p$.

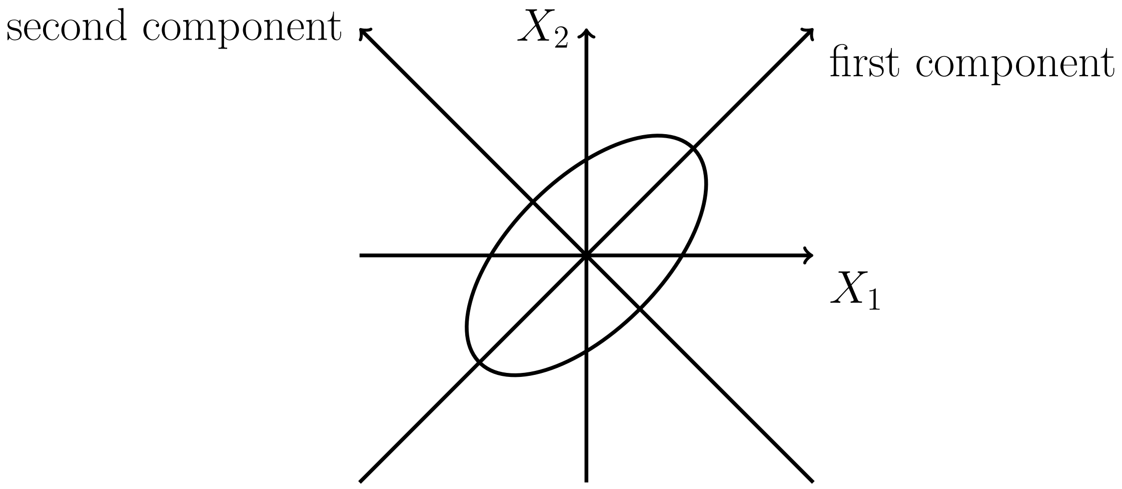
Note that for an arbitrary matrix, $\mathbf{B}_{p \times q}$, the $\mathrm{Tr}(\mathbf{B}\mathbf{B}')$ = sum of squares of the entries in $\mathbf{B}$. Also, for two matrices, $\mathbf{B}$ and $\mathbf{C}$, if both of the products $\mathbf{BC}$ and $\mathbf{CB}$ can be taken, then $\mathrm{Tr}(\mathbf{BC})$ is equal to $\mathrm{Tr}(\mathbf{CB})$. Using these two results, the least-squares criterion value can be given as

$$\mathrm{Tr}([\lambda_{r+1} \mathbf{a}_{r+1} \mathbf{a}'_{r+1} + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}'_p][\lambda_{r+1} \mathbf{a}_{r+1} \mathbf{a}'_{r+1} + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}'_p]') =$$

$$\sum_{k \geq r+1} \lambda_k^2 \ .$$

This measure is one of how bad the rank r approximation might be (i.e., the proportion of unexplained sum-of-squares when put over $\Sigma_{k=1}^{p} \lambda_k^2$).

For a geometric interpretation of principal components, suppose we have two variables, X_1 and X_2, that are centered at their respective means (i.e., the means of the scores on X_1 and X_2 are zero). In the diagram below, the ellipse represents the scatter diagram of the sample points. The first principal component is a line through the widest part; the second component is the line at right angles to the first principal component. In other words, the first principal component goes through the fattest part of the "football", and the second principal component through the next fattest part of the "football" and orthogonal to the first; and so on. Or, we take our original frame of reference and do a rigid transformation around the origin to get a new set of axes; the origin is given by the sample means (of zero) on the two X_1 and X_2 variables. (To make these same geometric points, we could have used a constant density contour for a bivariate normal pair of random variables, X_1 and X_2, with zero mean vector.)



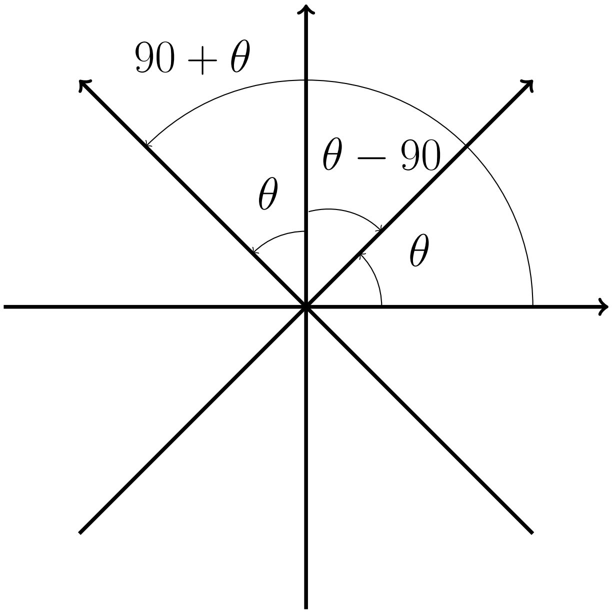
As an example of how to find the placement of the components in the picture given above, suppose we have the two variables, X_1 and X_2, with variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} .$$

Let a_{11} and a_{21} denote the weights from the first eigenvector of $\mathbf{\Sigma}$; a_{12} and a_{22} are the weights from the second eigenvector. If these are placed in a 2×2 orthogonal (or rotation) matrix $\mathbf{T}$, with the first column containing the first eigenvector weights and the second column the second eigenvector weights, we can obtain the direction cosines of the new axes system from the following:

$$\mathbf{T} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \cos(90+\theta) \\ \cos(\theta - 90) & \cos(\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} .$$

These are the cosines of the angles with the positive (horizontal and vertical) axes. If we wish to change the orientation of a transformed axis (i.e., to make the arrow go in the other direction), we merely use a multiplication of the relevant eigenvector values by -1 (i.e., we choose the other normalized eigenvector for that same eigenvalue, which still has unit length).



If we denote the data matrix in this simple two variable problem as $\mathbf{X}_{n \times 2}$, where n is the number of subjects and the two columns represent the values on variables X_1 and X_2 (i.e., the coordinates of each subject on the original axes), the $n \times 2$ matrix of coordinates of the subjects on the transformed axes, say $\mathbf{X}_{trans}$ can be given as $\mathbf{XT}$.

For another interpretation of principal components, the first component could be obtained by minimizing the sum of squared perpendicular residuals from a line (and in analogy to simple regression where the sum of squared vertical residuals from a line is minimized). This notion generalizes to more than than one principal component and in analogy to the way that multiple regression generalizes simple regression — vertical residuals to hyperplanes are used in multiple regression, and perpendicular residuals to hyperplanes are used in PCA.

There are a number of specially patterned matrices that have interesting eigenvector/eigenvalue decompositions. For example, for the $p \times p$ diagonal variance-covariance matrix

$$\Sigma_{p\times p} = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_p^2 \end{pmatrix} ,$$

the roots are $\sigma_1^2, \ldots, \sigma_p^2$, and the eigenvector corresponding to σ_i^2 is $[0\ 0 \ldots 1 \ldots 0]'$ where the single 1 is in the i^{th} position. If we have a correlation matrix, the root of 1 has multiplicity p, and the eigenvectors could also be chosen as these same vectors having all zeros except for a single 1 in the i^{th} position, $1 \leq i \leq p$.

If the $p \times p$ variance-covariance matrix demonstrates compound symmetry,

$$\Sigma_{p\times p} = \sigma^2 \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & & \vdots \\ \rho & \cdots & 1 \end{pmatrix} ,$$

or is an equicorrelation matrix,

$$\mathbf{P} = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & & \vdots \\ \rho & \cdots & 1 \end{pmatrix} ,$$

then the $p - 1$ smallest roots are all equal. For example, for the equicorrelation matrix, $\lambda_1 = 1 + (p - 1)\rho$, and the remaining $p - 1$ roots are all equal to $1 - \rho$. The $p \times 1$ eigenvector for λ_1 is $[\frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}]'$, and defines an average. Generally, for any variance-covariance matrix with all entries greater than zero (or just non-negative), the entries in the first eigenvector must all be greater than zero (or non-negative). This is known as the Perron-Frobenius theorem.

Although we will not give these tests explicitly here (they can be found in Johnson and Wichern's (2007) multivariate text), they are inference methods to test the null hypothesis of an equicorrelation matrix (i.e., the last $p-1$ eigenvalues are equal); that the variance-covariance matrix is diagonal or the correlation matrix is the identity (i.e., all eigenvalues are equal); or a sphericity test of independence that all eigenvalues are equal and $\mathbf{\Sigma}$ is σ^2 times the identity matrix.

0.1 Analytic Rotation Methods

Suppose we have a $p \times m$ matrix, $\mathbf{A}$, containing the correlations (loadings) between our p variables and the first m principal components. We are seeking an orthogonal $m \times m$ matrix $\mathbf{T}$ that defines a rotation of the m components into a new $p \times m$ matrix, $\mathbf{B}$, that contains the correlations (loadings) between the p variables and the newly rotated axes: $\mathbf{AT} = \mathbf{B}$. A rotation matrix $\mathbf{T}$ is sought that produces "nice" properties in $\mathbf{B}$, e.g., a "simple structure", where generally the loadings are positive and either close to 1.0 or to 0.0.

The most common strategy is due to Kaiser, and calls for maximizing the normal varimax criterion:

$$\frac{1}{p}\sum_{j=1}^{m}[\sum_{i=1}^{p}(b_{ij}/h_i)^4 - \frac{\gamma}{p}\{\sum_{i=1}^{p}(b_{ij}/h_i)^2\}^2] \; ,$$

where the parameter $\gamma = 1$ for varimax, and $h_i = \sqrt{\sum_{j=1}^{m} b_{ij}^2}$ (this is called the square root of the communality of the i^{th} variable in a factor analytic context). Other criteria have been suggested for this so-called orthomax criterion that use different values of γ — 0 for quartimax, $m/2$ for equamax, and $p(m-1)/(p+m-2)$ for parsimax. Also, various methods are available for attempting oblique rotations where the transformed axes do not need to maintain orthogonality, e.g., oblimin in SYSTAT; Procrustes in MATLAB.

Generally, varimax seems to be a good default choice. It tends to "smear" the variance explained across the transformed axes rather evenly. We will stick with varimax in the various examples we do later.

0.2 Little Jiffy

Chester Harris named a procedure posed by Henry Kaiser for “factor analysis”, Little Jiffy. It is defined very simply as “principal components (of a correlation matrix) with associated eigenvalues greater than 1.0 followed by normal varimax rotation”. To this date, it is the most used approach to do a factor analysis, and could be called “the principal component solution to the factor analytic model”.

More explicitly, we start with the $p \times p$ sample correlation matrix $\mathbf{R}$ and assume it has r eigenvalues greater than 1.0. $\mathbf{R}$ is then approximated by a rank r matrix of the form:

$$\begin{gathered}
\mathbf{R} = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \cdots + \lambda_r \mathbf{a}_r \mathbf{a}_r' = \\
(\sqrt{\lambda_1}\mathbf{a}_1)(\sqrt{\lambda_1}\mathbf{a}_1') + \cdots + (\sqrt{\lambda_r}\mathbf{a}_r)(\sqrt{\lambda_r}\mathbf{a}_r') = \\
\mathbf{b}_1 \mathbf{b}_1' + \cdots + \mathbf{b}_r \mathbf{b}_r' = \\
(\mathbf{b}_1, \ldots, \mathbf{b}_r) \begin{pmatrix} \mathbf{b}_1' \\ \vdots \\ \mathbf{b}_r' \end{pmatrix} = \mathbf{B}\mathbf{B}' \ ,
\end{gathered}$$

where

$$\mathbf{B}_{p \times r} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1r} \\ b_{21} & b_{22} & \cdots & b_{2r} \\ \vdots & \vdots & & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pr} \end{pmatrix} \ .$$

The entries in $\mathbf{B}$ are the loadings of the row variables on the column components.

For any $r \times r$ orthogonal matrix $\mathbf{T}$, we know $\mathbf{T}\mathbf{T}' = \mathbf{I}$, and

$$\mathbf{R} = \mathbf{B}\mathbf{I}\mathbf{B}' = \mathbf{B}\mathbf{T}\mathbf{T}'\mathbf{B}' = (\mathbf{B}\mathbf{T})(\mathbf{B}\mathbf{T})' = \mathbf{B}^*_{p \times r}\mathbf{B}^{*'}_{r \times p} \ .$$

For example, varimax is one method for constructing $\mathbf{B}^*$. The columns of $\mathbf{B}^*$ when normalized to unit length, define r linear composites of the observable variables, where the sum of squares within columns of $\mathbf{B}^*$ defines the variance for that composite. The composites are still orthogonal.

0.3 Principal Components in Terms of the Data Matrix

For convenience, suppose we transform our $n \times p$ data matrix $\mathbf{X}$ into the z-score data matrix $\mathbf{Z}$, and assuming $n > p$, let the SVD of $\mathbf{Z}_{n\times p} = \mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}'_{p\times p}$. Note that the $p \times p$ correlation matrix

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}'\mathbf{Z} = \frac{1}{n}(\mathbf{VDU}')(\mathbf{UDV}') = \mathbf{V}(\frac{1}{n}\mathbf{D}^2)\mathbf{V}' \; .$$

So, the rows of $\mathbf{V}'$ are the principal component weights. Also,

$$\mathbf{ZV} = \mathbf{UDV}'\mathbf{V} = \mathbf{UD} \; .$$

In other words, $(\mathbf{UD})_{n\times p}$ are the scores for the n subjects on the p principal components.

What's going on in "variable" space: Suppose we look at a rank 2 approximation of $\mathbf{Z}_{n\times p} \approx \mathbf{U}_{n\times 2}\mathbf{D}_{2\times 2}\mathbf{V}'_{2\times p}$. The i^{th} subject's row data vector sits somewhere in p-dimensional "variable" space; it is approximated by a linear combination of the two eigenvectors (which gives another point in p dimensions), where the weights used in the linear combination come from the i^{th} row of $(\mathbf{UD})_{n\times 2}$. Because we do least-squares, we are minimizing the squared Euclidean distances between the subject's row vector and the vector defined by the particular linear combination of the two eigenvectors. These approximating vectors in p dimensions are all in a plane defined by all linear combinations of the two eigenvectors. For a rank 1 approximation, we merely have a multiple of the first eigenvector (in p dimensions) as the approximating vector for a subject's row vector.

What's going on in "subject space": Suppose we begin by looking at a rank 1 approximation of $\mathbf{Z}_{n \times p} \approx \mathbf{U}_{n \times 1}\mathbf{D}_{1 \times 1}\mathbf{V}'_{1 \times p}$. The j^{th} column (i.e., variable) of $\mathbf{Z}$ is a point in n-dimensional "subject space", and is approximated by a multiple of the scores on the first component, $(\mathbf{UD})_{n \times 1}$. The multiple used is the j^{th} element of the $1 \times p$ vector of first component weights, $\mathbf{V}'_{1 \times p}$. Thus, each column of the $n \times p$ approximating matrix, $\mathbf{U}_{n \times 1}\mathbf{D}_{1 \times 1}\mathbf{V}'_{1 \times p}$, is a multiple of the same vector giving the scores on the first component. In other words, we represent each column (variable) by a multiple of one specific vector, where the multiple represents where the projection lies on this one single vector (the term "projection" is used because of the least-squares property of the approximation). For a rank 2 approximation, each column variable in $\mathbf{Z}$ is represented by a point in the plane defined by all linear combinations of the two component score columns in $\mathbf{U}_{n \times 2}\mathbf{D}_{2 \times 2}$; the point in that plane is determined by the weights in the j^{th} column of $\mathbf{V}'_{2 \times p}$. Alternatively, $\mathbf{Z}$ is approximated by the sum of two $n \times p$ matrices defined by columns being multiples of the first or second component scores.

As a way of illustrating a graphical way of representing principal components of a data matrix (through a biplot), suppose we have the rank 2 approximation, $\mathbf{Z}_{n \times p} \approx \mathbf{U}_{n \times 2}\mathbf{D}_{2 \times 2}\mathbf{V}'_{2 \times p}$, and consider a two-dimensional Cartesian system where the horizontal axis corresponds to the first component and the vertical axis corresponds to the second component. Use the n two-dimensional coordinates in $(\mathbf{U}_{n\times 2}\mathbf{D}_{2\times 2})_{n\times 2}$ to plot the rows (subjects), let $\mathbf{V}_{p\times 2}$ define the two-dimensional coordinates for the p variables in this same space. As in any biplot, if a vector is drawn from the origin through the i^{th} row (subject) point, and the p column points are projected onto this vector, the collection of such projections is proportional to the i^{th} row of the $n \times p$ approximation matrix $(\mathbf{U}_{n\times 2}\mathbf{D}_{2\times 2}\mathbf{V}'_{2\times p})_{n\times p}$.

The emphasis in this notes has been on the descriptive aspects of principal components. For a discussion of the statistical properties of these entities, consult Johnson and Wichern (2007) — confidence intervals on the population eigenvalues; testing equality of eigenvalues; assessing the patterning present in an eigenvector; and so on.