What Is Principle Components Analsysis Anyway?

Jeff Gill, jgill@polisci.ufl.edu

1 Principal Components Analysis

This note briefly explains the theory and practice of principal components analysis. The basic idea is to linearly transform a dataset into smaller dimension dataset with the property that each of the transformed variables is uncorrelated. This process, originally from Pearson (1901) and Hotelling (1933), is designed to reveal structure in the data that would otherwise be difficult to observe. Recent detailed descriptions include Flury (1988), Jackson (1991), Krzanowski (1988),

1.1 Introduction

Principal components analysis (PCA) is a means of transforming data through rotation in the sample space of observations. Each variable defines a dimension and therefore an axis. PCA re-expresses the variability of the data such that the total amount of variance is preserved but:

- Axes are enumerated in descending order of variance explained. That is, the first dimension explains the most variance, the second dimension explains the second-most variance, and so on.
- The new axes are uncorrelated with each other: they are orthogonal.
- If there exists correlation in the original data then it is expressed as zero length along some dimensions after the rotation of axes.

The last point is not always clear. If there are p variables in the original data and there is some correlation between these variables, then PCA produces a rotation in the p dimensions except that some number of these, q, will be of zero length, where the magnitude of q indicates the extent of the correlation between variables and $p - q$ indicates the extent of orthogonal information in the data.

Therefore if all p variables are uncorrelated then the axes are already orthogonal and there is no need to perform PCA. Conversely, if the p variables are perfectly correlated then there exists only one dimension worth of information in the data and all but one of the axes will have data of zero-length after PCA.

Consider for a moment only two variables: X_1 and X_2, which are assumed for simplicity to have mean zero each and unit variance. The correlation between any two variables is defined as:

$$\rho = \frac{COV(X_1, X_2)}{\sqrt{VAR(X_1)VAR(X_2)}},$$

where VAR and COV indicate the variance and covariance respectively. This formula obviously reduces to the covariance since the variances are assumed to be equal to one. If the correlation between these two variables is actually zero, then the equiprobability contours (concentric lines indicating equal probability of occurrence) of these two variables is circular. On the other hand, if there is a non-zero ρ value then the shape of the equiprobability contours will be elliptical where the cosine of the angle of intersection from the longest elliptical axis to the original x-axis (measured at the origin since zero mean is assumed for both variables) is equal to ρ. In the extreme case of perfect correlation between X_1 and X_2 the equiprobability contours condense to a single line.

1.2 The Theory

Begin with an $n \times p$ data matrix $\mathbf{X}$, where variables are organized in columns, and standardize. Define $\mathbf{R}$ as the correlation matrix corresponding to $\mathbf{X}$ along with a matrix $\mathbf{E}$ of the eigenvectors of the $\mathbf{R}$ matrix with the constraint that squared rows and columns of $\mathbf{E}$ sum to one. Then by standard spectral theory (Lax 1997, Chapter 6), the matrix defined by:

$$\boldsymbol{\lambda} = \mathbf{E}'\mathbf{R}\mathbf{E} \tag{2}$$

is a matrix containing the descending eigenvalues along the diagonal and zeros elsewhere. The structure of this diagonal matrix is revealing. It is in fact the variance-covariance matrix of the rotation defined by the principal components. So each eigenvalue, $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$, is the variance of a principal component where the first principal component now accounts for the largest variance by construction. Furthermore, since the off-diagonal elements are all zero, then the correlation has been removed in the new coordinate system. It is typical for some proportion of the lower diagonal elements of $\boldsymbol{\lambda}$ to be zero or within machine epsilon of zero, and the number of non-zero elements indicates the rank of the $\mathbf{X}$ matrix.

The importance of the $\mathbf{E}$ matrix is that it provides the transformation of the data points from the original metric to the PCA metric through simple matrix multiplication:

$$\mathbf{Y} = \mathbf{X}\mathbf{E}. \tag{3}$$

Thus $\mathbf{Y}$ are the points in the new rotated coordinate system where the variance structure is preserved: the *principal component scores*. The usefulness of this transformation is that it is one-to-one and therefore reversible, i.e.

$$\mathbf{X} = \mathbf{E}'\mathbf{Y} \tag{4}$$

because of the orthogonal property of the $\mathbf{E}$ matrix described below.

The $\mathbf{E}$ matrix of normalized eigenvectors also has some interesting properties. It is orthogonal, meaning that:

$$\mathbf{E}'\mathbf{E} = \mathbf{E}\mathbf{E}' = \mathbf{I}. \tag{5}$$

This property also allows us to modify (1) according to:

$$\begin{aligned}\boldsymbol{\lambda}\mathbf{E}' &= \mathbf{E}'\mathbf{R}\mathbf{E}\mathbf{E}' \\ \boldsymbol{\lambda}\mathbf{E}' &= \mathbf{E}'\mathbf{R} \\ 0 &= |\mathbf{R} - \boldsymbol{\lambda}|.\end{aligned}$$

The last expression is called the characteristic equation of the $\mathbf{R}$ matrix.

We can also define a new matrix according to:

$$\mathbf{L} = \mathbf{E}\boldsymbol{\lambda}^{\frac{1}{2}} \tag{6}$$

where the square root on $\boldsymbol{\lambda}$ is simply the square root of each diagonal element. This matrix square root calculation is trivial for two reasons: the diagonality of the matrix means that there is no need for a Cholesky decomposition, and the fact that the diagonal values are guaranteed to be positive means that complex roots are avoided. The $\mathbf{L}$ matrix is theoretically important due to two related multiplicative properties:

$$\begin{aligned}\mathbf{L}\mathbf{L}' &= \mathbf{E}\boldsymbol{\lambda}^{\frac{1}{2}}(\mathbf{E}\boldsymbol{\lambda}^{\frac{1}{2}})' \\ &= \mathbf{E}\boldsymbol{\lambda}\mathbf{E}' \\ &= \mathbf{E}(\mathbf{E}'\mathbf{R}\mathbf{E})\mathbf{E}' \\ &= \mathbf{R}\end{aligned}$$

and:

$$\begin{aligned}\mathbf{L}'\mathbf{L} &= (\mathbf{E}\boldsymbol{\lambda}^{\frac{1}{2}})'\mathbf{E}\boldsymbol{\lambda}^{\frac{1}{2}} \\ &= (\boldsymbol{\lambda}^{\frac{1}{2}})'\mathbf{E}'\mathbf{E}\boldsymbol{\lambda}^{\frac{1}{2}} \\ &= \boldsymbol{\lambda}\end{aligned} \tag{7}$$

So the product of component loadings is either equal to the correlation matrix ($\mathbf{R}$) or the diagonal eigenvalue matrix ($\boldsymbol{\lambda}$), depending on the order of matrix multiplication.

1.3 A Worked Example

Begin with the contrived by illustrative data matrix given by:

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 1 & 2 \\ 4 & 3 & 6 \\ 5 & 5 & 5 \\ 6 & 7 & 6 \\ 7 & 9 & 9 \\ 8 & 8 & 8 \\ 9 & 8 & 3 \end{bmatrix}.$$

Now standardize the column variables to produce:

$$\mathbf{X} = \begin{bmatrix} -1.461 & -1.109 & -1.385 \\ -1.095 & -0.421 & -0.652 \\ -0.730 & -1.453 & -1.018 \\ -0.365 & -0.765 & 0.448 \\ 0.000 & -0.076 & 0.081 \\ 0.365 & 0.612 & 0.448 \\ 0.730 & 1.300 & 1.547 \\ 1.095 & 0.956 & 1.181 \\ 1.461 & 0.956 & -0.652 \end{bmatrix}.$$

The correlation matrix from $\mathbf{X}$ is:

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.880 & 0.619 \\ 0.880 & 1.000 & 0.716 \\ 0.619 & 0.716 & 1.000 \end{bmatrix}.$$

The eigenvalues and eigenvectors are found by solving the characteristic equation: $|\mathbf{R} - \boldsymbol{\lambda}| = 0$. This produces the matrices:

$$\mathbf{E} = \begin{bmatrix} 0.585 & -0.514 & 0.628 \\ 0.607 & -0.236 & -0.759 \\ 0.538 & 0.825 & 0.174 \end{bmatrix}, \qquad \boldsymbol{\lambda} = \begin{bmatrix} 2.482 & 0.00 & 0.000 \\ 0.000 & 0.41 & 0.000 \\ 0.000 & 0.00 & 0.108 \end{bmatrix}.$$

This means that the proportion of the total variance explained by the first principal component explains $2.482/3 = 0.827$. By the same reasoning, the second principal component explains 13.7% of the total variance and the third principal component explains 3.6% of the total variance.

The component scores are produced by pre-multiplying the original data matrix by $\mathbf{E}$. This produces:

$$\mathbf{Y} = \begin{bmatrix} -2.272 & -0.130 & -0.316 \\ -1.247 & 0.124 & -0.482 \\ -1.857 & -0.122 & 0.467 \\ -0.437 & 0.737 & 0.429 \\ -0.003 & 0.085 & 0.072 \\ 0.826 & 0.038 & -0.157 \\ 2.049 & 0.595 & -0.259 \\ 1.856 & 0.186 & 0.168 \\ 1.084 & -1.513 & 0.078 \end{bmatrix}.$$

Here the mean for each $\mathbf{Y}$ variable (the columns) remains zero and the corresponding variance is no longer unity but rather the corresponding diagonal value of $\boldsymbol{\lambda}$. Because of these component scores result from the simple matrix multiplication defined in (2),

they are in fact linear combinations of the original data with weights determined by the eigenvector matrix **E**. For instance, the first value of **Y** is produced from:

$$\begin{aligned} Y_{11} &= \sum_{j=1}^{3} E_{.1} X_{1.} \\ &= 0.585 \times -1.461 + 0.607 \times -1.109 + 0.538 \times -1.385 \\ &= -2.272. \end{aligned}$$

1.4 Discussion

The primary purpose of the PCA process is to reveal interrelationships within the data through the transformation to the PCA metric. Some of these characteristics are obvious. If two variables are perfectly correlated (one is a linear combination of the other) then their corresponding correlation with the other variables is identical and this shows up as an all-zero row and column of the $\boldsymbol{\lambda}$ matrix. Obviously then the determinant of the $\boldsymbol{\lambda}$ matrix is then zero and this forms a quick check on the system. If the eigenvalues are identical then the marginal probability contours are circular and the complete joint probability contours are spherical in p-space. Thus the eigenvalues immediately provide information about interrelationships between variables.

References

Flury, Bernhard. 1988. *Common Principal Components and Related Multivariate Models.* New York: Wiley & Sons.

Hotelling, H. 1933. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Education Psychology* 24, 417-441.

Jackson, J. Edward. 1991. *A User's Guide to Principal Components.* New York: Wiley & Sons.

Jolliffe, I. T. 1986. *Principal Components Analysis.* New York: Springer-Verlag.

Krzanowski, W. J. 1988. *Principles of Multivariate Analysis.* Oxford: Oxford University Press.

Lax, Peter D. 1997. *Linear Algebra.* New York: Wiley & Sons.

Pearson, Karl. 1901. "On Lines and Planes of Closest fit to Systems of Points in Space." *Philosophical Magazine* Series 6, 2, 559-72.