Principal Components Analysis

The objective in a *principal components analysis* (or *PCA*) of a set of variables is to find a new set of variables that have certain desirable properties, and that efficiently represent the information in the original set of variables (efficiently here means with a smaller number of new variables than the original set.)

The numerical analysis problem that must be solved in PCA is that given p variables, $X_1, X_2, \ldots, X_p$, find the linear combinations (new variables) $Z_1, Z_2, \ldots, Z_p$, e.g.

$$Z_{1i} = a_{11}X_{1i} + a_{12}X_{2i} + \cdots + a_{1p}X_{pi}, \ i = 1, \ldots, n,$$

that have the following properties:

    a) $\operatorname{var}(Z_1) \geq \operatorname{var}(Z_2) \geq \ldots \geq \operatorname{var}(Z_p)$,

    b) the **a**'s should have the maximum simultaneous resemblance to all X's, and

    c) $\operatorname{cov}(Z_j, Z_k) = 0$, all $k \neq j$.

Usually, but not always, the X's have been rescaled to have a mean of 0 and a standard deviation of 1 (i.e. transformed to *z-scores*.

The first linear combination, Z_1, looks somewhat like a regression equation, and the properties are designed so that a) the (new) components are defined in decreasing order of importance, b) the components represent the "underlying dimensions" of the set of original variables, and c) are not correlated with one another.

The first, or principal, component (hence the name of the analysis) is

$$Z_{1i} = a_{11}X_{1i} + a_{12}X_{2i} + \cdots + a_{1p}X_{pi}, \ i = 1, \ldots, n,$$

and is defined by choosing the a's in order to

$$\max(\lambda_1) = \operatorname{var}(Z_1) = \frac{1}{n}\sum_{i=1}^{n} Z_{1i}^2,$$

subject to the constraint that $a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1$. This constraint ensures that λ_1 is not made arbitrarily large simply by increasing the values of the a's.

A second component

$$Z_{2i} = a_{21}X_{1i} + a_{22}X_{2i} + \cdots + a_{2p}X_{pi}, \ i = 1, \ldots, n,$$

can also be defined in order to

$$\max(\lambda_2) = \operatorname{var}(Z_2) = \frac{1}{n}\sum_{i=1}^{n} Z_{2i}^2,$$

subject to the constraints that $a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1$, and $\mathrm{cov}(Z_1, Z_2) = 0$. These constraints again make sure that λ_2 is not made arbitrarily large, and that the first and second components are independent or uncorrelated with one another.

The optimization problem amounts to requiring the first component to be defined in such a way as to have the maximum *variance* over the n observations. The optimization problem may be stated as:

$$\max(\lambda_1) = \frac{1}{n}\sum_{i=1}^{n} Z_{1i}^2, \text{ s.t. } \sum_{i=1}^{n} a_{1i}^2 = 1$$

$$= \frac{1}{n}(\mathbf{a}_1'\mathbf{X})^2, \text{ s.t. } \mathbf{a}_1'\mathbf{a}_1 = 1$$

$$= \frac{1}{n}(\mathbf{a}_1'\mathbf{X})(\mathbf{X}'\mathbf{a}_1), \text{ s.t. } \mathbf{a}_1'\mathbf{a}_1 = 1$$

$$= \mathbf{a}_1'(\frac{1}{n}\mathbf{X}\mathbf{X}')\mathbf{a}_1, \text{ s.t. } \mathbf{a}_1'\mathbf{a}_1 = 1.$$

Note that if $\mu_X = 0$, and $\sigma_X = 1$, (i.e. that the X's are z-scores), $\frac{1}{n}\mathbf{X}\mathbf{X}' = \mathbf{R}$, the correlation matrix.

Maximizing λ_1, i.e.

$$\max(\lambda_1) = \mathbf{a}_1'\mathbf{R}\mathbf{a}_1, \text{ s.t. } \mathbf{a}_1'\mathbf{a}_1 = 1$$

is equivalent to maximizing u_1 u_1, i.e.

$$\max(u_1) = \mathbf{a}_1'\mathbf{R}\mathbf{a}_1 - \lambda_1(\mathbf{a}_1'\mathbf{a}_1 - 1),$$

where λ_1 appears in the equation as a "Lagrange multiplier" which incorporates the constraint into the equation.

To maximize u_1, set the partial derivative of u with respect to $\mathbf{a}$ to zero, i.e.,

$$\frac{\partial u_1}{\partial \mathbf{a}} = 2\mathbf{R}\mathbf{a} - 2\lambda_1\mathbf{a}_1 = \mathbf{0}, \text{ or}$$

$$(\mathbf{R} - \lambda_1)\mathbf{a}_1 = \mathbf{0}.$$

This last expression contains p simultaneous equations that must be solved for $\mathbf{a}_1$.

Comparison of this expression with that describing the "eigenstructure" of a square matrix reveals that choosing $\mathbf{a}_1$ to be the first eigenvector of $\mathbf{R}$, with λ_1 as the first eigenvalue, solves the optimization problem in PCA. It also happens that the second and higher-numbered components are described by the second and higher-numbered eigenvectors and eigenvalues of the matrix $\mathbf{R}$.

There is a second way to state the optimization problem that winds up leading to the same result. In this second case, instead of choosing the coefficients, $\mathbf{a}_1$, such that the first component has the maximum variability over the observations, instead $\mathbf{a}_1$ is defined to have the maximum simultaneous resemblance to the X's, i.e.,

$$\max(\lambda_1) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}(a_{1j}X_{ji})^2 / \sum_{j=1}^{p}a_{1j}^2\right)$$

$$= \frac{1}{n}(\mathbf{a}_1'\mathbf{X})^2 / (\mathbf{a}'\mathbf{a})$$

$$= \frac{1}{n}(\mathbf{a}_1'\mathbf{X}\mathbf{X}'\mathbf{a}_1) / (\mathbf{a}'\mathbf{a})$$

$$= (\mathbf{a}_1'\mathbf{R}'\mathbf{a}_1) / (\mathbf{a}'\mathbf{a}),$$

which leads to the same problem as before, i.e.

$$\max(\lambda_1) = \mathbf{a}_1'\mathbf{R}\mathbf{a}_1, \text{ s.t. } \mathbf{a}_1'\mathbf{a}_1 = 1.$$