

Economics 310
Handout #III

Principal Components

Let \mathbf{X} be an $n \times k$ data matrix where each variable is measured as deviations from its own mean. The matrix $\mathbf{X}'\mathbf{X}$ consists of sums of squares and cross products of the various variables which make up the data matrix. $\text{Trace}(\mathbf{X}'\mathbf{X}) =$ the total variation in the data set.

The first principal component of \mathbf{X} is a vector $\mathbf{z}_1 = \alpha_{11}\mathbf{x}_1 + \alpha_{12}\mathbf{x}_2 + \dots + \alpha_{1k}\mathbf{x}_k$, a linear combination of the column vectors of the \mathbf{X} matrix, which minimizes the sum of the squared distances (orthogonal to \mathbf{z}_1) from the various vectors in \mathbf{X} to \mathbf{z}_1 . However since any if any \mathbf{z}_1 has this property any other vector $k\mathbf{z}_1$ where k is any scalar has this property also. So we need another restriction on \mathbf{a}_1 . The restriction we choose is that $\mathbf{a}_1'\mathbf{a}_1 = \mathbf{1}$. It can readily be seen that \mathbf{z}_1 maximizes the sum of the squared lengths of the projections of the vectors which make up the \mathbf{X} matrix onto \mathbf{z}_1 subject to the same restriction. Then our problem is to find the vector \mathbf{z}_1 which maximizes the sum $\widehat{\mathbf{x}}_1'\widehat{\mathbf{x}}_1 + \widehat{\mathbf{x}}_2'\widehat{\mathbf{x}}_2 + \dots + \widehat{\mathbf{x}}_k'\widehat{\mathbf{x}}_k$ where $\widehat{\mathbf{x}}_i$ is

the fitted vector from the regression of \mathbf{x}_i on \mathbf{z}_1 subject to the restriction that $\sum_{i=1}^k a_{ii}^2 = \mathbf{a}_1'\mathbf{a}_1 = \mathbf{1}$.

(This restriction assures that $\mathbf{z}_1'\mathbf{z}_1 = \widehat{\mathbf{x}}_1'\widehat{\mathbf{x}}_1 + \widehat{\mathbf{x}}_2'\widehat{\mathbf{x}}_2 + \dots + \widehat{\mathbf{x}}_k'\widehat{\mathbf{x}}_k$.) Formally then we maximize

$\mathbf{z}_1'\mathbf{z}_1$ subject to the condition that $\mathbf{a}_1'\mathbf{a}_1 = \mathbf{1}$ where $\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$. In a similar way we can find the second principal component which is the vector $\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2$ which maximizes $\mathbf{z}_2'\mathbf{z}_2$ subject to the normalization $\mathbf{a}_2'\mathbf{a}_2 = \mathbf{1}$ and subject to the additional restriction that the second principal component is orthogonal to the first.

Find the first principal component of a data set X:

Choose the vector \mathbf{a}_1 to maximize $\mathbf{z}_1'\mathbf{z}_1 = (\mathbf{X}\mathbf{a}_1)'(\mathbf{X}\mathbf{a}_1)$ subject to $\mathbf{a}_1'\mathbf{a}_1 = \mathbf{1}$.

The lagrangian is

$$L = \mathbf{a}_1'\mathbf{X}'\mathbf{X}\mathbf{a}_1 + \lambda_1(1 - \mathbf{a}_1'\mathbf{a}_1)$$

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\mathbf{X}'\mathbf{X}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = \mathbf{0}$$

This yields the equation $[\mathbf{X}'\mathbf{X} - \lambda_1\mathbf{I}]\mathbf{a}_1 = \mathbf{0}$. You will immediately recognize that this is an eigenvalue problem.. The solution is that λ_1 is the largest eigenvalue of $\mathbf{X}'\mathbf{X}$ and \mathbf{a}_1 is the associated eigenvector. The second principal component is $\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2$ where \mathbf{a}_2 is the eigenvector associated with the second largest eigenvalue and so forth until all k principal components have been found. Let \mathbf{Z} be the matrix each of whose columns is a principal component of \mathbf{X} . Then the total variation in $\mathbf{X} =$ the total variation in $\mathbf{Z} = \text{Trace}(\mathbf{X}'\mathbf{X}) = \text{Trace}(\mathbf{Z}'\mathbf{Z}) = \sum \lambda_i$ and the number of nonzero eigenvalues = the rank of the matrix \mathbf{X} . λ_i = the squared length of the i th principal component.