Lecture 3 and 4 Sept. 18 and Sept.20-2006

**Data Visualization**

STAT 442 / 890, CM 462

Lecture: Ali Ghodsi

# 1 Principal Components Analysis

Principal components analysis (PCA) is a very popular technique for dimensionality reduction. Given a set of data on $n$ dimensions, PCA aims to find a linear subspace of dimension $d$ lower than $n$ such that the data points lie mainly on this linear subspace (See Figure 1 as an example of a two-dimensional projection found by PCA). In practice we are not able to find a reduced subspace where all of the points lie exactly in that subspace. Instead we try to find a subspace which attempts to maintain most of the variability of the data.

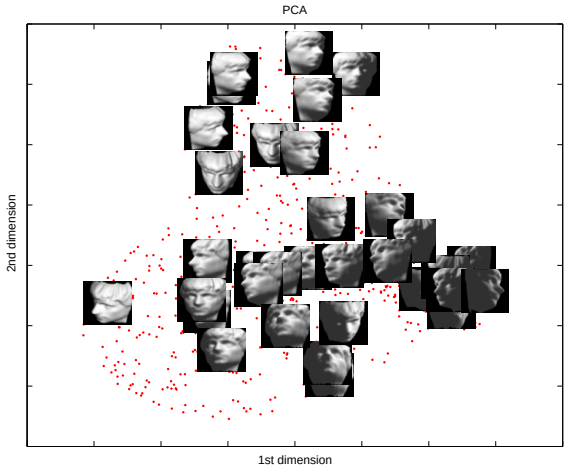


Figure 1: *PCA applied to the same data set. A two-dimensional projection is shown, with a sample of the original input images.*

The linear subspace can be specified by $d$ orthogonal vectors, call them: $U_1, U_2, \ldots, U_d$ that form a new coordinate system, called the 'principal components'. The principal components are orthogonal, linear transformations of the original data points, so there can be no more than $n$ of them. However, the hope is that only $d < n$ principal components are needed to approximate the space spanned by the $n$ original axes. In the case where $d = n$ the number of dimensions remains the same and there is no reduction.

**Example 1:** The following is a reduction from a two dimensional space into a one dimensional space. We can see the principal components $U_1$ and $U_2$ in the diagram. $U_1$ corresponds to the direction in which the data has the most variance while $U_2$ is orthogonal to it. If we ignore the $U_2$ direction and project the points in the $U_1$ dimension we reduced the dimensionality of the data.

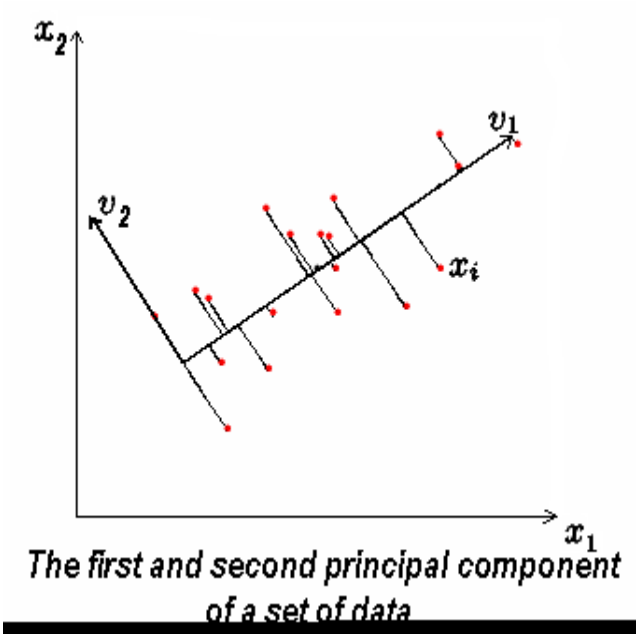


Figure 2: *PCA applied to a 2D data set to reduce it to a single dimension.*

**Example 2:** In this example we can reduce a set of 3D data to a set in only two dimensions. We do this by forming a plane from the first two principal components $U_1$ and $U_2$. We then project all of the points on to that plane. The points are shown in their reduced dimension in the figure on the right.

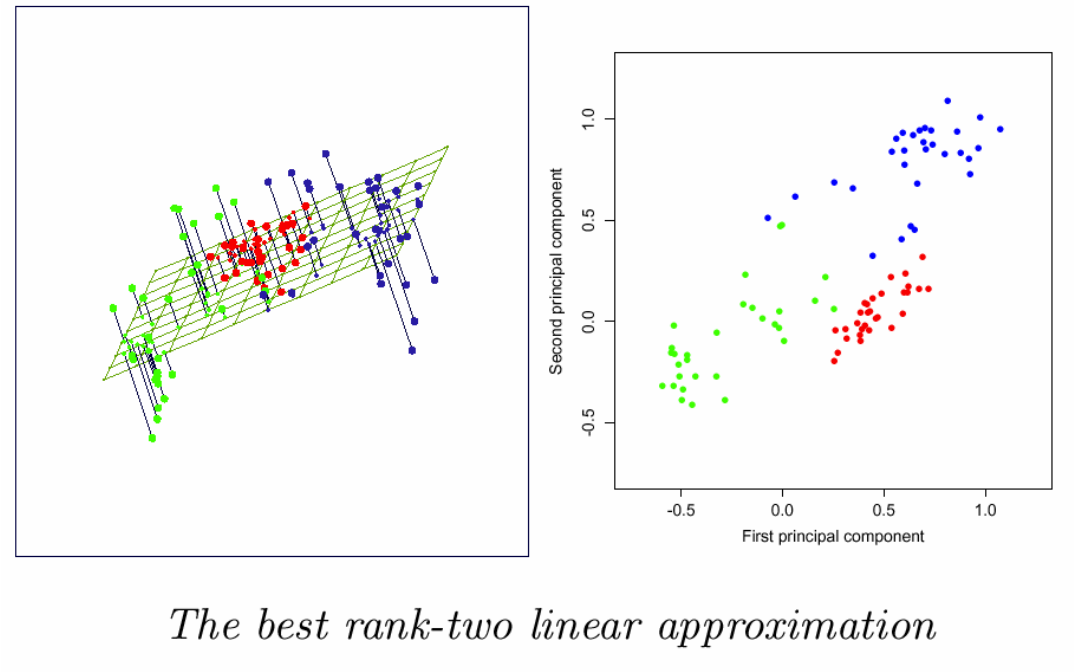


Figure 3: *PCA applied to a 3D data set to reduce it to two dimensions.*

The most common definition of PCA, due to Hotelling [1], is that, for a given set of data vectors $x_i$, $i \in 1...t$, the $d$ principal axes are those orthonormal axes onto which the variance retained under projection is maximal.

In order to capture as much of the variability as possible, let us choose the first principal component, denoted by $U_1$, to have maximum variance. Suppose that all centered observations are stacked into the columns of an $n \times t$ matrix $X$, where each column corresponds to

an $n$-dimensional observation and there are $t$ observations. Let the first principal component be a linear combination of $X$ defined by coefficients (or weights) $w = [w_1....w_n]$.

$$U_1 = w_1^{(1)}x_1 + w_2^{(1)}x_2 + \cdots + w_n^{(1)}x_n$$

In matrix form:

$$U_1 = w^T X$$

We want this first dimension to have maximum variance.

$$var(U_1) = var(w^T X) = w^T S w$$

where $S$ is the $n \times n$ sample covariance matrix of $X$.

Clearly $var(U_1)$ can be made arbitrarily large by increasing the magnitude of $w$. This means that the variance stated above has no upper limit and so we can not find the maximum. To solve this problem, we choose $w$ to maximize $w^T S w$ while constraining $w$ to have unit length. Therefore, we can rewrite the above equation as:

$$\max \ w^T S w$$

$$subject\ to\ w^T w = 1$$

To solve this optimization problem a Lagrange multiplier $\alpha_1$ is introduced:

$$L(w, \alpha) = w^T S w - \alpha_1(w^T w - 1) \tag{1}$$

Differentiating with respect to $w$ gives $n$ equations,

$$Sw = \alpha_1 w$$

Premultiplying both sides by $w^T$ we have:

$$w^T S w = \alpha_1 w^T w = \alpha_1$$

$var(U_1)$ is maximized if $\alpha_1$ is the largest eigenvalue of $S$.

Clearly $\alpha_1$ and $w$ are an eigenvalue and an eigenvector of $S$. Differentiating (1) with respect to the Lagrange multiplier $\alpha_1$ gives us back the constraint:

$$w^T w = 1$$

This shows that the first principal component is given by the normalized eigenvector with the largest associated eigenvalue of the sample covariance matrix $S$. A similar argument can show that the $d$ dominant eigenvectors of covariance matrix $S$ determine the first $d$ principal components.

# References

[1] H. Hotelling. Analysis of a complex of statistical variables into components. *J. of Educational Psychology*, 24:417–441, 1933.

**Algorithm 1**

**Recover basis:** Calculate $XX^\top = \sum_{i=1}^{t} x_i x_i^\top$ and let $U$ = eigenvectors of $XX^\top$ corresponding to the top $d$ eigenvalues.

**Encode training data:** $Y = U^\top X$ where $Y$ is a $d \times t$ matrix of encodings of the original data.

**Reconstruct training data:** $\hat{X} = UY = UU^\top X$.

**Encode test example:** $y = U^\top x$ where $y$ is a $d$-dimensional encoding of $x$.

**Reconstruct test example:** $\hat{x} = Uy = UU^\top x$.

Table 1: *Direct PCA Algorithm*