# Principal Components Analysis

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

August 22, 2008

Reference: Much of the material in this note is from Krzanowski, W. J. (1988). *Principles of Multivariate Analysis.* Oxford, UK: Oxford University Press.

**Summary**: This note starts by summarizing some of the main points regarding principal component analysis (PCA). A fuller discussion of these points is provided in the remainder of the note.

Consider a set of $N$ $d$-dimensional data vectors $\{\mathbf{x}^i\}_{i=1}^N$. We want to re-represent these data points using a new basis (and possibly in a $q$-dimensional space where $q < d$).

*Implementation*: Let

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^i \tag{1}$$

denote the data sample mean vector. Then the data sample covariance matrix $S$ is

$$S = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}^i - \overline{\mathbf{x}})\ (\mathbf{x}^i - \overline{\mathbf{x}})^T. \tag{2}$$

The new basis is formed by the eigenvectors of $S$. If dimensionality reduction is desired, ignore the eigenvectors with small eigenvalues.

PCA is generally derived in two ways, one way is based on finding directions in which the "variance of the data is maximized", and the other way is based on "minimizing reconstruction error". We next briefly describe these two ways of deriving PCA, and later show that they are equivalent to each other.

*Derivation of PCA I*: For a set of $d$-dimensional data vectors $\{\mathbf{x}^i\}_{i=1}^N$, the principal axes $\{\mathbf{e}\}_{j=1}^q$ are those orthonormal axes onto which the retained variance under projection is maximal. It can be shown that the vectors $\mathbf{e}_j$ are given by the $q$ dominant eigenvectors of the sample covariance matrix $S$, such that $S\mathbf{e}_j = \lambda_j \mathbf{e}_j$. The $q$ principal components of the observed vector $\mathbf{x}^i$ are given by the vector

$$\mathbf{t}^i = E^T (\mathbf{x}^i - \overline{\mathbf{x}}) \tag{3}$$

where $E$ is a matrix whose $j^{\text{th}}$ column is $\mathbf{e}_j$. The variables $t_j$ (the elements of the vector $\mathbf{t}$) are then uncorrelated such that the covariance matrix $\sum_i (\mathbf{t}^i)(\mathbf{t}^i)^T / N$ is diagonal with elements $\lambda_j$.

*Derivation of PCA II*: Of all orthogonal linear projections $\mathbf{t}^i = E^T(\mathbf{x}^i - \overline{\mathbf{x}})$, the principal component projection minimizes the squared reconstruction error $\sum_i \|\mathbf{x}^i - \hat{\mathbf{x}}^i\|^2$ where the optimal linear reconstruction of $\mathbf{x}^i$ is given by $\hat{\mathbf{x}}^i = E^T\mathbf{t}^i + \overline{\mathbf{x}}$.

**Fuller discussion**: Consider an imaginary data set in which the heights ($X_1$) and weights ($X_2$) of $n$ individuals have been recorded. This two-dimensional sample might then be represented (after mean-centering) by a scatter plot such as that in Figure 1.

The axis $OX_1$ and $OX_2$ ($O$ stands for origin) of this representation have been determined by the variables height and weight respectively. Note, however, that we could rotate the axes to the new positions $OY_1$ and $OY_2$ without altering the configuration of points, and relate the points to these new axes for any future analysis without changing the outcome of that analysis.

It may also occur that such new axes may actually carry some useful meaning to the investigator; indeed they may even be more meaningful than the original measurements that were taken. For instance, imagine all the data points successively compressed on to the $OY_1$ and $OY_2$ axes. Points at the extreme right-hand end of $OY_1$ will correspond to individuals that have large values of both height and weight, whereas points at the extreme left of $OY_1$ have small values of both height and weight. Thus an individual's value on the $OY_1$ axis is a reflection of that individual's size, and hence this axis can be labeled as a 'size' axis. Now, consider axis $OY_2$. Points at the top of this axis will tend to correspond to individuals whose weight ($X_2$) is large in relation to their height ($X_1$), whereas points at the bottom of this axis will tend to correspond to individuals whose height is large in relation to their weight. Thus an individual's value on the $OY_2$ axis is a reflection of that individual's shape, and this axis can be labeled as a 'shape' axis.

Axes $OY_1$ and $OY_2$ are a rotation of axes $OX_1$ and $OX_2$, and so the relationship between the two representations is

$$\begin{aligned} y_1 &= x_1 \cos\alpha + x_2 \sin\alpha && (4) \\ y_2 &= -x_1 \sin\alpha + x_2 \cos\alpha && (5) \end{aligned}$$

where $\alpha$ is the angle between $OX_1$ and $OY_1$ or between $OX_2$ and $OY_2$. For a fixed value of $\alpha$, this is a linear relationship.

In Figure 1 there is a wide spread of sample values on the $OY_1$ axis, and a relatively small spread of values on the $OY_2$ axis. This means that individuals have widely differing sizes, but similar shapes. It is tempting therefore to approximate the two-dimensional data by a one-dimensional approximation to this data that is obtained by projecting the points on to the $OY_1$ axis. This will give a reasonably good approximation since we could characterize the differences between the $n$ individuals sufficiently well if, instead of quoting the height $x_1$ and weight $x_2$ for each, we were simply to quote its index of size $y_1$. Replacing the two original variables $X_1$ and $X_2$ by a single derived variable $Y_1$ effects a reduction in dimensionality from 2 to 1.

Different values of the angle $\alpha$ give different derived variables $Y_1$. Among the different possible values, there will be one that is deemed to be 'best.' Consider the point $P_i$ in Figure 1, and its projection onto the $OY_1$ axis labeled $P_1'$. The line between $P_i$ and $P_i'$ is denoted $P_iP_i'$, and its length is the displacement of the point from its two-dimensional representation to its one-dimensional

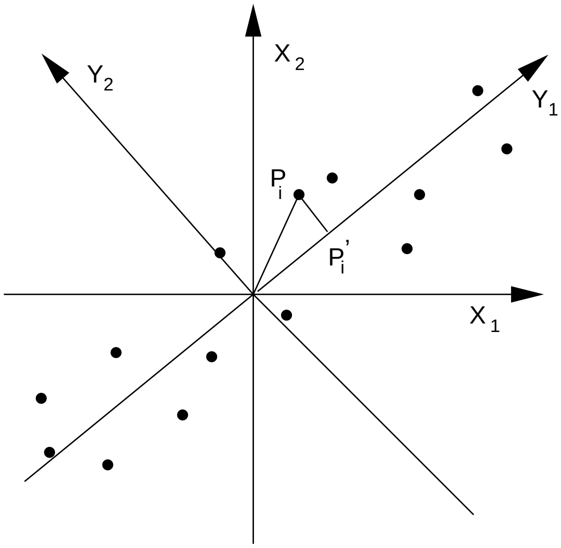

Figure 1: Sample of heights ($X_1$) and weights ($X_2$) of $n$ individuals.

representation. We will regard the optimal value of $\alpha$ as the value that gives rise to the smallest displacement of all the points from their original two-dimensional positions. In other words, the line $OY_1$ of closest fit to the points is defined to be the one obtained by minimizing $\sum_{i=1}^{n}(P_iP_i')^2$.

Applying Pythagoras' Theorem to the triangle $OP_iP_i'$, we get

$$(OP_i)^2 = (OP_i')^2 + (P_iP_i')^2. \tag{6}$$

Summing over all the points $P_i$, it follows

$$\sum_{i=1}^{n}(OP_i)^2 = \sum_{i=1}^{n}(OP_i')^2 + \sum_{i=1}^{n}(P_iP_i')^2. \tag{7}$$

Hence

$$\frac{1}{n-1}\sum_{i=1}^{n}(OP_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(OP_i')^2 + \frac{1}{n-1}\sum_{i=1}^{n}(P_iP_i')^2. \tag{8}$$

The left-hand side of this equation is fixed for any given sample regardless of the coordinate-system that is used. Hence choosing $OY_1$ to minimize $\frac{1}{n-1}\sum_{i=1}^{n}(P_iP_i')^2$ is equivalent to choosing $OY_1$ to maximize $\frac{1}{n-1}\sum_{i=1}^{n}(OP_i')^2$. Since $O$ is the centroid of the points (recall that the data is mean-centered), $\frac{1}{n-1}\sum_{i=1}^{n}(OP_i')^2$ is just the sample variance when the individuals have values given by their $Y_1$ coordinate. Thus finding the line $OY_1$ that minimizes the sum of squared perpendicular deviations of the points from this line is exactly equivalent to finding the line $OY_1$ such that the projections of points on it have maximum variance.

It'll now be useful if we abandon our simple two-dimensional example, and instead consider the more general case when data lies in a $p$-dimensional space. In what follows below, we use a sequence of steps of the above form. The data are modeled as usual by a swarm of $n$ points in $p$

dimensions, each axis corresponding to a measured variable. First, we look for a line $OY_1$ in this space such that the spread of the $n$ points when projected on to this line is a maximum. Having obtained $OY_1$, we next consider the $(p-1)$ dimensional subspace orthogonal to $OY_1$, and look for the line $OY_2$ in this subspace such that the spread of points when projected on to this line is a maximum. This is equivalent to seeking a line $OY_2$ at right angles to $OY_1$, such that the spread of points when they are projected onto $OY_2$ is as large as possible (although, clearly, this spread must be no greater than the spread along $OY_1$). Having, obtained $OY_1$ and $OY_2$, we then consider the $(p-2)$ dimensional subspace orthogonal to both $OY_1$ and $OY_2$. Thus we look for a line $OY_3$ which is at right angles to both $OY_1$ and $OY_2$ such that the spread of points when projected along $OY_3$ is as large as possible after the spread on $OY_1$ and $OY_2$ have been taken into account. This process can be continued until we have obtained $p$ mutually orthogonal lines $OY_i (i = 1, \ldots, p)$.

The above discussion should give you an intuitive feel for principal components analysis. We are now about to dive into the mathematical details. Before doing so, we need to take two detours: one detour on sample covariance matrices and one detour on Lagrange multipliers.

We suppose that each data item is a $p$ dimensional vector $x_i = [x_{i1}, \ldots, x_{ip}]^T$, and that there are $n$ data items. These items can be placed in an $(n \times p)$ data matrix (each row is an individual data item) whose $(i, j)$th element is $x_{ij}$. The mean (across all data items) of the $j$th variable is $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. Let all these means be collected together in the sample mean vector $\bar{\mathbf{x}} = [\bar{x}_1, \ldots, \bar{x}_p]^T$. The variance of the $j$th variable is given by

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \tag{9}$$

and the covariance between the $j$th and $k$th variables is given by

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k). \tag{10}$$

Let all these variances and covariances be collected together in the sample covariance matrix $\mathbf{S}$ which has $(j, k)$th element $s_{jk}$. By expanding the right-hand side as a matrix product, it can be shown that

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \tag{11}$$

Now consider forming a variable $Y$ as a linear combination $a_1X_1 + \ldots + a_pX_p$ of the original variables $X_i$. We can more concisely write $Y = \mathbf{a}^T\mathbf{X}$ where $\mathbf{a} = [a_1, \ldots, a_p]^T$. Then the value of $Y$ corresponding to the $i$th data item is

$$\begin{aligned} y_i &= a_1x_{i1} + \ldots + a_px_{ip} & (12) \\ &= \mathbf{a}^T\mathbf{x}_i, & (13) \end{aligned}$$

and the mean of $Y$ over the $n$ sample members is

$$\begin{aligned} \bar{y} &= a_1\bar{x}_1 + \ldots + a_p\bar{x}_p & (14) \\ &= \mathbf{a}^T\bar{\mathbf{x}}. & (15) \end{aligned}$$

Next consider the sample variance of $Y$. This is given by

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{16}$$

Because

$$\begin{aligned} y_i - \bar{y} &= \mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\bar{\mathbf{x}} && (17)\\ &= \mathbf{a}^T(\mathbf{x}_i - \bar{\mathbf{x}}) && (18)\\ &= (\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{a}, && (19) \end{aligned}$$

then

$$(y_i - \bar{y})^2 = \mathbf{a}^T(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T\mathbf{a}, \tag{20}$$

and we can write

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\mathbf{a}^T(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T\mathbf{a} && (21)\\ &= \mathbf{a}^T\{\sum_{i=1}^{n}(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T\}\mathbf{a}. && (22) \end{aligned}$$

Consequently,

$$\begin{aligned} S_Y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 && (23)\\ &= \mathbf{a}^T\{\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T\}\mathbf{a} && (24)\\ &= \mathbf{a}^T\mathbf{S}\mathbf{a}. && (25) \end{aligned}$$

If needed, it is easy to show that $\mathbf{a}^T\mathbf{S}\mathbf{a} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j s_{ij}$. We are now done with the detour on sample covariance matrices. We have shown that the variance of $Y$ can be expressed in terms of the covariance matrix of $X$, denoted $\mathbf{S}$, and the linear coefficients, denoted $\mathbf{a}$.

The next detour has to do with Lagrange optimization. This is a method for maximizing (or minimizing) a function subject to one or more constraints. We give an example of Lagrange optimization. From this example, you should get the general idea of how this method works.

Example: Suppose that we want to find the shortest vector $\mathbf{y} = [y_1\ y_2]^T$ that lies on the line $2y_1 - y_2 - 5 = 0$. We write down an objective function $L$ in which the first term states the function that we want to minimize, and the second term states the constraint:

$$L = \frac{1}{2}(y_1^2 + y_2^2) + \lambda(2y_1 - y_2 - 5) \tag{26}$$

where $\lambda$ is called a Lagrange multiplier (it is always placed in front of the constraint). The first term gives the length of the vector (actually its 1/2 times the length of the vector squared); the second term gives the constraint. We now take derivatives, and set these derivatives equal to zero:

$$\frac{\partial L}{\partial y_1} = y_1 + 2\lambda = 0 \tag{27}$$

$$\frac{\partial L}{\partial y_2} = y_2 - \lambda = 0 \tag{28}$$

$$\frac{\partial L}{\partial \lambda} = 2y_1 - y_2 - 5 = 0. \tag{29}$$

Note that we have three equations and three unknown variables. From the first derivative we know that $y_1 = -2\lambda$, and from the second derivative we know that $y_2 = \lambda$. If we plug these values into the third derivative, we get $-4\lambda - \lambda = 5$ or that $\lambda = -1$. From this, we can now solve for $y_1$ and $y_2$; in particular, $y_1 = 2$ and $y_2 = -1$. So the solution to our problem is $\mathbf{y} = [2 \ -1]^T$.

We are now done with detours, and can return to the problem of deriving principal components. Based on our discussion above, we define the first principal component to be the linear combination $Y_1 = \mathbf{a}_1^T\mathbf{X}$ of the original variables such that the variance of $Y_1$, $\mathbf{a}_1^T\mathbf{S}\mathbf{a}_1$, is maximal. Note, however, that there is a problem. One uninteresting way of making $Y_1$ have a large variance is by making the linear coefficients $\mathbf{a}_1 = [a_{11}, \ldots, a_{1p}]^T$ be as large as possible (i.e. infinite). This would be a trivial solution. To avoid this solution, we must constrain the coefficients so that they are bounded. In particular, we use the constraint that the sum of squares of the coefficients is equal to one, i.e.

$$\mathbf{a}_1^T\mathbf{a}_1 = \sum_{i=1}^{p} a_{1i}^2 = 1. \tag{30}$$

Let

$$V_1 = \mathbf{a}_1^T\mathbf{S}\mathbf{a}_1 - \lambda_1(\mathbf{a}_1^T\mathbf{a}_1 - 1) \tag{31}$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p} a_{1i}a_{1j}s_{ij} - \lambda_1(\sum_{i=1}^{p} a_{1i}^2 - 1) \tag{32}$$

Then

$$\frac{\partial V_1}{\partial a_{1k}} = 2\sum_{j=1}^{p} s_{kj}a_{1j} - 2\lambda_1 a_{1k} \quad (k = 1, \ldots, p) \tag{33}$$

To find the vector $\mathbf{a}_1 = [a_{11}, \ldots, a_{1p}]^T$ maximizing $V_1$, we set $\frac{\partial V_1}{\partial a_{1k}} = 0$ for all $k$ and solve the resulting set of equations. For the $k$th coefficient we get

$$\sum_{j=1}^{p} s_{kj}a_{1j} = \lambda_1 a_{1k}. \tag{34}$$

The left-hand side of this equation is the $k$th element of $\mathbf{S}\mathbf{a}_1$, while the right-hand side is the $k$th element of $\lambda_1\mathbf{a}_1$. Thus when all $k$ equations are treated simultaneously, it follows that the maximizing value of $\mathbf{a}_1$ must satisfy

$$\mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1. \tag{35}$$

That is, the vector of coefficients $\mathbf{a}_1$ is an eigenvector of the data covariance matrix $\mathbf{S}$ with eigenvalue $\lambda_1$. We choose $\mathbf{a}_1$ to be the eigenvector with the largest eigenvalue.

We are done! I will not bore you with the details proving that the linear coefficients corresponding to the other principal components are also given by the eigenvectors of $\mathbf{S}$. Suffice it to say that one uses Lagrange optimization where we maximize the variance of $Y_i$ subject to the constraints that $\mathbf{a}_i^T \mathbf{a}_i = 1$ and also that $\mathbf{a}_i$ is orthogonal to all previous vectors of linear coefficients. The fact that all our intuitions about good coordinate systems and good dimensionality reduction turns out to be an eigenvalue/eigenvector problem is an amazing result! (Are you amazed?)