PRINCIPAL COMPONENT ANALYSIS

1 INTRODUCTION

One of the main problems inherent in statistics with more than two variables is the issue of visualising or interpreting data. Fortunately, quite often the problem can be simplified by replacing a group of variables with a single new variable. The reason might be that more than one variable is measuring the same driving principle governing the behaviour of the system. One of the methods for reducing variables is Principal Component Analysis (PCA).

The purpose of the report is to give precise mathematical definition to PCA. Then, mathematical derivation of PCA is given.

2 PRINCIPAL COMPONENT ANALYSIS

The method creates a new set of variables called principal components. Each of the new variables is a linear combination of the original variables. Each of principal components is chosen so that it would describe most of the still available variance and all principal components are orthogonal to each other; hence there is no redundant information. The first principal component has the maximum variance among all possible choices. (The MathWorks, 2010) (Jolliffe, 1986)

PCA is used for different purposes - finding interrelations between variables in the data; interpreting and visualizing data; decreasing the number of variables for making further analysis simpler and for many other similar reasons.

The definition and derivation of the principal component analysis is described. In between Lagrange multipliers for finding a maximum of a function with constraints and eigenvalues and eigenvectors are explained, because these ideas are needed in the derivation.

2.1 DEFINITION OF PRINCIPAL COMPONENTS

Suppose that x is a vector of r random variables and x^T denotes the transpose of x. So

$$x = [x_1, x_2, \dots, x_r\ ]^T$$

First step is to look at the linear function $\alpha_1^T x$ of the elements of x which has maximum variance, where α_1 is a vector of r constants, $\propto_{11}, \propto_{12}, \dots, \propto_{1r}$, so that

$$\alpha_1^T x = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1r}x_r = \sum_{j=1}^{r} \alpha_{1j}x_j$$

There must be some constraints imposed, otherwise variance is unbounded. In the current paper $\alpha_1^T \alpha_1 = 1$ is used, which means that the sum of squares of elements of α_1 is 1 or the length of α_1 is 1.

So, the aim is to find the linear function that transforms random variables into a new random variable so that the new variable $\alpha_1^T x$ has maximum variation.

Next, it is necessary to find a linear function $\alpha_2^T x$, uncorrelated with $\alpha_1^T x$, which has maximum variance, and then for linear function $\alpha_3^T x$, uncorrelated with $\alpha_1^T x \; and \; \alpha_2^T x$ and so on up to r^{th} linear function such that k^{th} linear function $\alpha_k^T x$ is uncorrelated with $\alpha_1^T x, \alpha_2^T x, \dots, \alpha_{k-1}^T x$.

All of these transformations create *r* new random variables which are called principal components. In general, the hope is that the first few random variables are needed to explain necessary amount of variability in the dataset.

2.1.1 EXAMPLE

In the simplest case, there is a pair of 2 random variables X and Y, which are highly correlated. In this case one might want to reduce the variables to only one, so that it would be easier to conduct further analysis. Figure 1 shows 30 such pairs of $(X_i, Y_i) \; for \; i = 1,2, \dots ,30$ random variables, where X_i is a random number from interval $(-1,1)$ and Y_i is $0.6X_i$ plus 0.4 times a random number from interval $(-1,1)$.

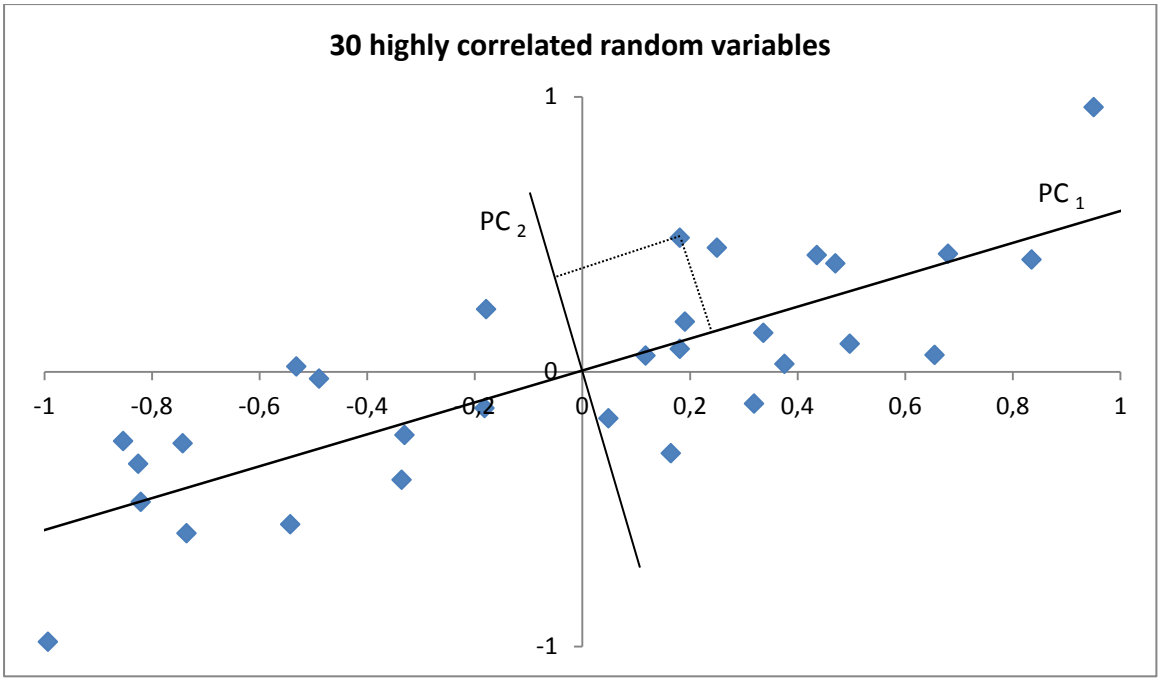

Figure 1 – 30 highly correlated random variables. Source: Random sample

A strong correlation is visible. Principal component analysis tries to find the first principal component which would explain most of the variance in the dataset. In this case it is clear that the most variance would stay present if the new random variable (first principal component) would be on the direction shown with the line on the graph. This new random variable would explain most of the variation in the data set and could be used for further analysis instead the original variables.

The method with two random variables looks similar to regression model, but the difference is that the first principle component is chosen so that the sample points are as close to the new variable as possible, but in regression analysis the vertical distances are as small as possible.

In reality, random variables X and Y can have some meaning as well. For example, X might be standardized mathematics exam scores and Y might be standardized physics exam scores. In that case it would be possible to conclude that the new variable PC_1 (the first principal component) might account for some general logical ability and PC_2 could be interpreted as some other factor.

2.2 DERIVATION OF PRINCIPAL COMPONENTS

The following part shows how to find those principal components. Basic structure of the definition and derivation are from I. T. Jolliffe's (1986) book "Principal Component Analysis".

It is assumed that the covariance matrix of the random variables x is known – denoted $\mathbb{M}$. $\mathbb{M}$ is a non-singular symmetric matrix with dimension r**.** $\mathbb{M}$ is also positive semi-definite which means that all the eigenvalues are non-negative. The element (i, j) of the matrix $\mathbb{M}$ shows the covariance between x_i and x_j in case $j \neq i$. Elements (i, i) on the diagonal show the variance of the element x_i. So,

$$\mathbb{M} = \begin{bmatrix} E[(x_1 - \mu_1)(x_1 - \mu_1)] & E[(x_1 - \mu_1)(x_2 - \mu_2)] & \cdots & E[(x_1 - \mu_1)(x_r - \mu_r)] \\ E[(x_2 - \mu_2)(x_1 - \mu_1)] & E[(x_2 - \mu_2)(x_2 - \mu_2)] & \cdots & E[(x_2 - \mu_2)(x_r - \mu_r)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(x_r - \mu_r)(x_1 - \mu_1)] & E[(x_r - \mu_r)(x_2 - \mu_2)] & \cdots & E[(x_r - \mu_r)(x_r - \mu_r)] \end{bmatrix},$$

where $E[x_i]$ is the expected value of x_i and μ_i is the mean of x_i. In this report the mean is assumed to be 0 because it can be subtracted from the data before the analysis.

Finding the principal components is reduced to finding eigenvalues and eigenvectors of the matrix $\mathbb{M}$ such that the k^{th} principal component is given by $z_k = \alpha_k^T x$. Here α_k is an eigenvector of $\mathbb{M}$, which corresponds to the k^{th} largest eigenvalue λ_k. In addition to this, the variance of z_k is λ_k because α_k is chosen to be unit length. (Jolliffe, 1986)

Before the result is derived two topics must be explained. Firstly, eigenvalues and eigenvector are described together with an example, and then method of Lagrange Multipliers is explained.

2.2.1 EIGENVALUES AND EIGENVECTORS

Eigenvector is a non-zero vector that stays parallel after matrix multiplication, i.e. x is eigenvector of dimension r of matrix $\mathbb{M}$ with dimension $r \times r$ if $\mathbb{M}x$ and x are parallel. Parallel means that there exists λ such that $\mathbb{M}x = \lambda x$. (Roberts, 1985)

To find eigenvalues and eigenvectors the equation $\mathbb{M}x = \lambda x$ must be solved. Rewrite it $(\mathbb{M} - \lambda I)x = 0$, where x and λ are both unknowns. Therefore $(\mathbb{M} - \lambda I)$ must be a singular matrix for non trivial eigenvectors. So, it is possible to find all λ's first because it is known from linear algebra that determinant of a singular matrix is 0. This equation is called eigenvalue equation. (ibid)

In case of symmetric semi-definite matrix where $a_{ij} = a_{ji}$, eigenvalues are non-negative real numbers and more importantly the eigenvectors are perpendicular to each other (ibid).

After finding all the eigenvalues λ's, all the corresponding eigen vectors can be found by solving standard linear matrix equation $Ax = 0$, where $A = (\mathbb{M} - \lambda I)$.

EXAMPLE OF FINDING EIGENVECTORS AND EIGENVALUES

Q: Let $A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ be a matrix. We need to find its eigenvectors and eigenvalues. (Strang, 1999)

A:

$$\det(A - \lambda I) = 0$$

$$\begin{bmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{bmatrix} = 0$$

$$(3-\lambda)^2 - 1 = 0$$

$$\lambda^2 - 6\lambda + 8 = 0$$

$$\lambda_1 = 4 \; and \; \lambda_2 = 2$$

Next, eigenvector for $\lambda_1 = 4$ can be found.

$$(A - 4I)x_1 = 0$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} x_1 = 0$$

$$x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

In a similar way for $\lambda_2 = 2$, $x_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Therefore, two pairs of eigenvalues and eigenvectors have been found as required.

In practice eigenvalues are computed by better algorithms than finding the roots of algebraic equations.

2.2.2 LAGRANGE MULTIPLIERS

Sometimes it is needed to find the maximum or minimum of the function that depends upon several variables whose values must satisfy certain equalities, i.e. constraints. In this report, it is needed to find principal components which are linear combination of original

random variables so that the length of the vector that represents linear combination is 1 and that all these vectors are uncorrelated to the others. The idea is to change the constrained n variable problem to unconstrained $n + 1$ variable problem (Gowers, 2008).

Situation can be stated as follows: maximize $f(x, y)$ subject to $g(x, y) = 0$ as illustrated in figures 1 and 2.

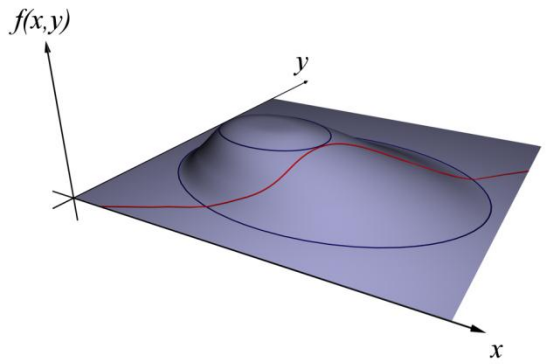

Figure 2 - Find *x* and *y* to maximize *f*(*x*,*y*) subject to a constraint (shown in red) *g*(*x*,*y*) = *c*. Source: (Lagrange multipliers)

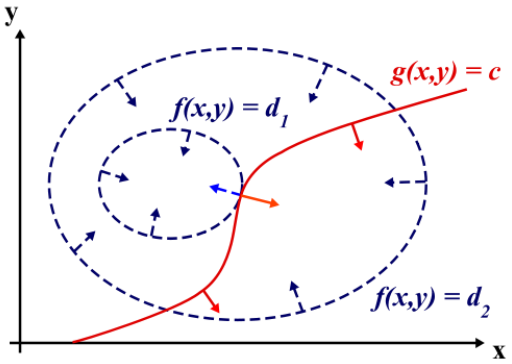

Figure 3 - Contour map of Figure 1. The red line shows the constraint *g*(*x*,*y*) = *c*. The blue lines are contours of *f*(*x*,*y*). The point where the red line tangentially touches a blue contour is the solution. Source: (Lagrange multipliers)

New variable λ called a Lagrange multiplier is introduced, and the Lagrange function is defined by

$$\Lambda(x, y, \lambda) = f(x, y) - \lambda g(x, y).$$

Needed extremum points are solutions of

$$\nabla_{xy\lambda}\Lambda(x, y, \lambda) = \left(\frac{\partial\Lambda}{\partial x}, \frac{\partial\Lambda}{\partial y}, \frac{\partial\Lambda}{\partial\lambda}\right) = 0.$$

Lagrange multiplier method gives necessary conditions for finding the maximum points of a function subject to constraints.

2.2.3 DERIVATION OF FIRST PRINCIPAL COMPONENT

It is needed to find α_1 (subject to $\alpha_1^T\alpha_1 = 1$ i.e. $\alpha_1^T\alpha_1 - 1 = 0$), which maximizes the variance of $\alpha_1^T x$. As $\mathbb{M}$ is the covariance matrix defined above and the data is normalized, i.e. $E[x] = 0$,

$$\alpha_1^T x = \sum_{i=1}^{r} \alpha_{1i} x_i$$

$$var(\alpha_1^T x) = E[(\alpha_1^T x)(\alpha_1^T x)] = E[(\alpha_1^T x)(\alpha_1^T x)^T] =$$

$$= E[\alpha_1^T x x^T \alpha_1] = \alpha_1^T E[x^T x]\alpha_1 = \alpha_1^T \mathbb{M}\alpha_1$$

The technique of Lagrange multipliers is used. So it is necessary to maximize

$$\alpha_1^T \mathbb{M}\alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1),$$

where λ is a Lagrange multiplier.

Differentiation with respect to α_1 gives

$$\mathbb{M}\alpha_1 - \lambda\alpha_1 = 0,$$

which is the same as

$$(\mathbb{M} - \lambda I_r)\alpha_1 = 0,$$

where I_r is the $(r \times r)$ identity matrix, i.e.

$$I_r = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Next, it is needed to decide which of the eigenvectors gives the maximizing value for the first principal component. It is necessary to maximize $\alpha_1^T \mathbb{M}\alpha_1$. Let α_1 be any eigenvector of $\mathbb{M}$ and λ be the corresponding eigenvalue. We have

$$\alpha_1^T \mathbb{M}\alpha_1 = \alpha_1^T \lambda\alpha_1 = \lambda\alpha_1^T \alpha_1 = \lambda$$

As λ must be the largest possible, α_1 must be the eigenvector which corresponds to the largest eigenvalue.

The first principal component has now been derived. Same process can be applied to others such that k^{th} principal component of x is $\alpha_k^T x$ and variance of $\alpha_k^T x$ is λ_k such that λ_k is the k^{th} largest eigenvalue with α_k the corresponding eigenvector of $\mathbb{M}$, where $k = 1,2, \dots, r$.

In this report, only the case where $k = 2$ will be proven.

The second principal component $\alpha_2^T x$ must maximize $\alpha_2^T \mathbb{M}\alpha_2$ such that $\alpha_2^T x$ is uncorrelated with $\alpha_1^T x$ i.e. covariance between $\alpha_2^T x$ and $\alpha_1^T x$ is 0. It can be written as

$$cov(\alpha_1^T x, \alpha_2^T x) = E[(\alpha_1^T x,)(\alpha_2^T x)^T] = \alpha_1^T E[x x^T]\alpha_2 = \alpha_1^T \mathbb{M}\alpha_2 =$$

$$= \alpha_2^T \lambda_1 \alpha_1^T = \lambda_1 \alpha_1^T \alpha_2 = \lambda_1 \alpha_2^T \alpha_1,$$

where $cov(x, y)$ denotes the covariance between x and y, and α_1 is already known from the derivation of first principal component. So, $\alpha_2^T \mathbb{M} \alpha_2$ must be maximized with following constraints: $\alpha_2^T \alpha_2 = 1 \; and \; \alpha_2^T \alpha_1 = 0$. The method of Lagrange multipliers is used.

$$\alpha_2^T \mathbb{M} \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \phi \alpha_2^T \alpha_1,$$

where $\lambda \; and \; \phi$ are Lagrange multipliers. As before, we differentiate with respect to α_2

$$\mathbb{M}\alpha_2 - \lambda\alpha_2 - \phi\alpha_1 = 0,$$

and for simplifying the left side it must be multiplied by α_1^T

$$\alpha_1^T \mathbb{M}\alpha_2 - \lambda\alpha_1^T\alpha_2 - \phi\alpha_1^T\alpha_1 = 0.$$

We know that $\alpha_1^T \mathbb{M}\alpha_2 = 0, \alpha_1^T\alpha_2 = 0 \; and \; \alpha_1^T\alpha_1 = 1$, so the equation reduces to $\phi = 0$. If we put $\phi = 0$ in the first equation then

$$\mathbb{M}\alpha_2 - \lambda\alpha_2 = 0,$$

$$(\mathbb{M} - \lambda I_p)\alpha_2 = 0,$$

where λ is an eigenvalue and α_k the corresponding eigenvector of $\mathbb{M}$. As before, $\alpha_1^T \mathbb{M}\alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda\alpha_1^T\alpha_1 = \lambda$, so $\lambda = \lambda_2 \; and \; \lambda \neq \lambda_1$ because λ must be as big as possible and due to correlations constraint it must not equal to λ_1.

As said before, the method applies for all the principal components and the proof is similar but not given in this report.

3 WORKS CITED

Gowers, T. (2008). *The Princeton Companion to Mathematics.* Princetone University Press.

Jolliffe, I. (1986). *Principal Component Analysis.* Harrisonburg: R. R. Donnelley & Sons.

Lagrange multipliers. (n.d.). Retrieved April 1, 2010, from Wikipedia: http://en.wikipedia.org/wiki/Lagrange_multipliers

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*(2), 559-572.

Roberts, A. W. (1985). *Elementary Linear Algebra.* Menlo Park: Benjamin/Cummings Pub. Co.

Strang, G. (1999). *MIT video lectures in linear algebra.* Retrieved April 16, 2010, from MIT Open Courseware: http://ocw.mit.edu/OcwWeb/Mathematics/18-06Spring-2005/VideoLectures/detail/lecture21.htm