

PRINCIPAL COMPONENTS ANALYSIS

Main Purpose and Intuition

Principal components analysis (PCA) is used for two objectives:

1. Reducing the number of variables comprising a dataset while retaining the variability in the data.
2. Identifying hidden patterns in the data, and classifying them according to how much of the information, stored in the data, they account for.

When mining a dataset comprised of numerous variables, (used interchangeably with the term dimensions hereinafter), it is likely that subsets of variables are highly correlated with each other. Given a high correlation between two or more variables it can be concluded that these variables are quite redundant thus share the same driving principle in defining the outcome of interest. In order to demonstrate this argument let us consider a basic example. Suppose we have measured 2 parametric properties (i.e. properties represented by numerical quantities) of a planar shape, which are the length and the width of the shape, that determine a certain outcome of interest [1]. From examining our observations we have noticed that these two properties seem to be positively correlated. Hence, we can replace them with a single new variable which is the area of the shape, that still captures most of the information about the shape supplied by its length and width.

In multivariate datasets, dimension reduction by PCA enables us to analyze our data in a visible 2-dimensional (2D) or 3D space, with only a mere loss of information.

The Basic Prerequisite - Correlation

Since PCA is mainly concerned with identifying correlations in the data, let us first focus our attention to the meaning of correlation. Correlation measures the simultaneous change in the values of two or more variables. There are numerous models for describing the behavioral nature of a simultaneous change in values, such as linear, exponential, periodic and more. The linear correlation is used in PCA.

The computational approach for defining correlation

Correlation between a pair of variables measures to what extent their values co-vary. The term covariance is undoubtedly associatively prompted immediately; not in vain, as covariance and correlation is nearly the same thing. The term in equation 1 is used to compute the covariance between a pair of variables (\bar{X}_1, \bar{X}_2) . (Actually, equation 1 computes the estimator for the covariance, as the values of the X variables are only a sample of the entire population).

$$COV(\bar{X}_1, \bar{X}_2) \equiv \frac{\sum_{i=1}^m (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{m}. \quad (1)$$

\bar{X} denotes the mean of \bar{X} and m denotes the number of points in each variable, i.e. the number of samples in each variable.

It can be concluded from equation 1 that the units covariance is measured by are the product of the units the variables are measured by. Consequently, the magnitude of the covariance is not easily interpretable, especially when the two variables are measured by different units.

Standardizing the values of the variables ($X_{ki}^s = (X_{ki} - \bar{X}_k) / \sigma_k$, where k denotes the variable index, i denotes the value index and s stands for standardized), gives the values in terms of standard deviation units from the variable's mean [2]. These units are known as *Z-scores*. Now each variable has the same mean (0), the same standard deviation (1), and each value is measured by the same terms (how much it deviates from the mean, in standard deviation units). Thus the values of the variables are now comparable.

Correlation between two variables (X_1, X_2) is measured by a term named *correlation coefficient* (denoted as r_{X_1, X_2} or simply as r) as presented in equation 2. This coefficient is also termed "Pearson product moment correlation coefficient".

$$r_{X_1, X_2} = \frac{\sum_{i=1}^m Z_{1i} Z_{2i}}{m} = \frac{\sum_{i=1}^m (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{m\sigma_1\sigma_2} = \frac{\sum_{i=1}^m (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^m (X_{1i} - \bar{X}_1)^2 (X_{2i} - \bar{X}_2)^2}} \quad (2)$$

σ denotes the standard deviation of \bar{X} .

The term in equation 2 is a reasonable way to measure correlation since paired *Z-scores* that do not vary in coordination from their means will contribute negatively to the sum in the numerator, whereas paired *Z-scores* that do vary in coordination from their means will contribute positively to the sum in the numerator. Thus, the sum of products discriminates between positive and negative co-variations, and measures the overall covariation. The average of this sum, which is computed by dividing it by the denominator in equation 2, gives a measure of the average covariation.

The visual approach for defining correlation

A visual aspect of correlation can be obtained by representing each one of a pair of variables as an axis in a Cartesian coordinate system, where the values of the variables are points plotted on the plane (figure 1). Correlation in this view, measures how well the model we believe describes the trend of the points in the plot, fits the actual trend in the plot.

It is extremely important to understand the visual aspects of correlation in order to properly use it. Such understanding may supply crucial insight about structures in the data being analyzed, and prevent potential biases that may arise by directly interpreting numerical results yielded by running computational procedures.

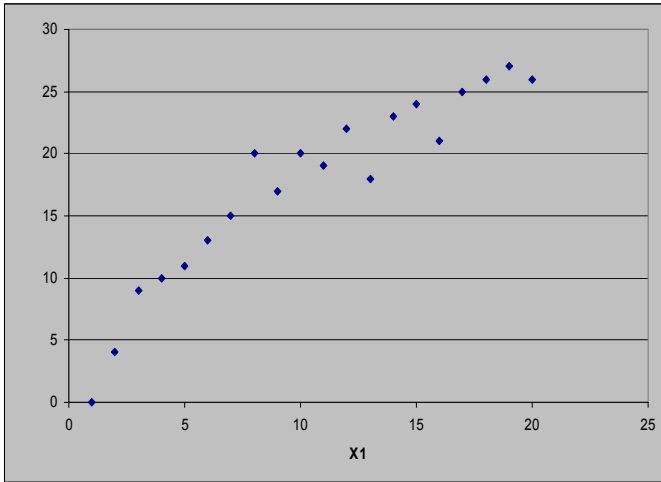


Figure 1: Scatter plot of a pair of variables.

Back to defining the visual meaning of correlation, it can be concluded from figure 1, that there is a connection between how close lays a putative trend line to the points in the plot, and how high the correlation is, under the model we believe best describes the actual trend in the data. Since PCA deals with linear correlation we will only focus on describing the connection between the points in the plot and a linear trend line. This linear trend line is constructed such that it minimizes the perpendicular distances, from it to each point in the plot. It is known as the least squares fitted line (figure 2).

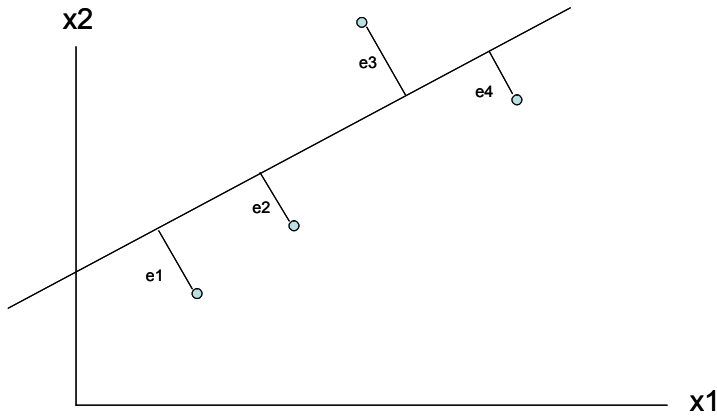


Figure 2: The best fit line is the one that minimizes the sum $e_1^2+e_2^2+e_3^2+e_4^2$. where e_i is the perpendicular distance from point i to the line. Such a line is the least square line (L.S. line).

Correlation might easily be confused with regression, thus it is important to stress out the differences between the two. In regression analysis the relation between the variable designated a criterion (dependent variable, i.e. Y) and the variable designated a predictor (independent variable, i.e. X) is tested. In this case an imperfect relation between Y and X is assumed, where Y contains an element of error (ϵ_i for each value $Y_i \in Y$), but the error in X is assumed to be negligible. Therefore, in order to regress Y upon X , X is treated as error-free and hence only the squared vertical distances are minimized. (See figure 3). However, correlation measures the symmetric relation between two random variables (therefore correlation $(X, Y) = \text{correlation}(Y, X)$), where

both X and Y contain amounts of error (δ_i for each value $X_i \in X$ and ε_i for each value $Y_i \in Y$, respectively) and hence the fitted correlation line is the line minimizing both vertical and horizontal distances [3] (see figure 3).

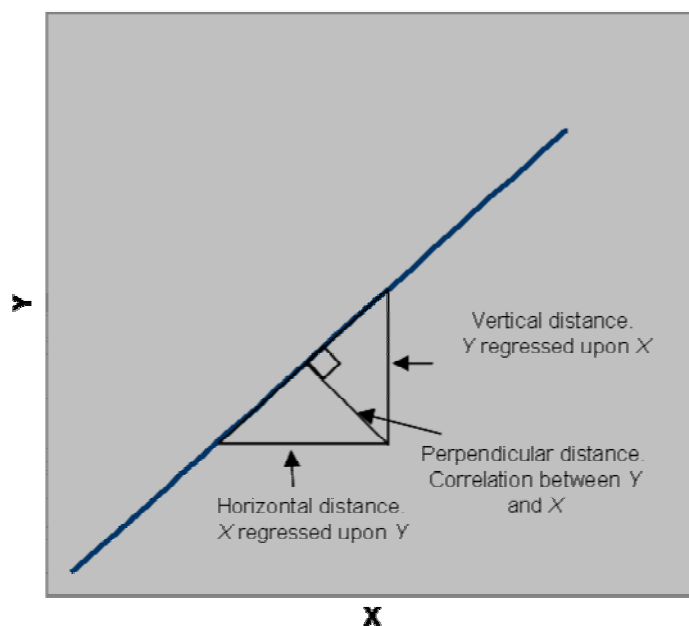


Figure 3: To regress Y upon X , the squared vertical distances are minimized; to regress X upon Y , the squared horizontal distances are minimized; to find the symmetric relation between X and Y , both vertical and horizontal distances are minimized.

The deviations of the points from the least squares fitted line are relative to the units the variables are measured by. Since different variables are measured by different units, their corresponding deviations would not be easily comparable or interpretable. The solution to this problem, developed for the Cartesian approach, is to use a ratio of some sort, which is unitless, to express the degree of fit of the model to the data.

The ratio used for expressing correlation is brought upon by comparing the deviations from the least squares fitted line to the deviations from a hypothetical line that is fitted to the same points as if they have no correlation. What would we require from such a line, measuring non-correlation? First, it should be horizontal (or vertical. The two cases are symmetric and so is the yielded ratio, thus only one is presented here). Any common variation between the two variables causes the least squares fitted line to angle up or down, depending on the nature of the correlation. Thus, forcing the comparative line to be horizontal prevents it from fitting to any sort of correlation. The next property of this line is its location on the axis. Using the least squares fitted line, we can predict the value of one of the variables based on the value of the other variable, as they share variation to a certain degree. In the case of the horizontal line, it represents a constant estimate: the mean of the variable whose value we are trying to predict. The intuition for using such an estimate is that in the case of complete uncertainty, the mean will always be the best guess [4]. Thus, this horizontal line should be located at the mean of the variable represented by the vertical axis (see figure 4).

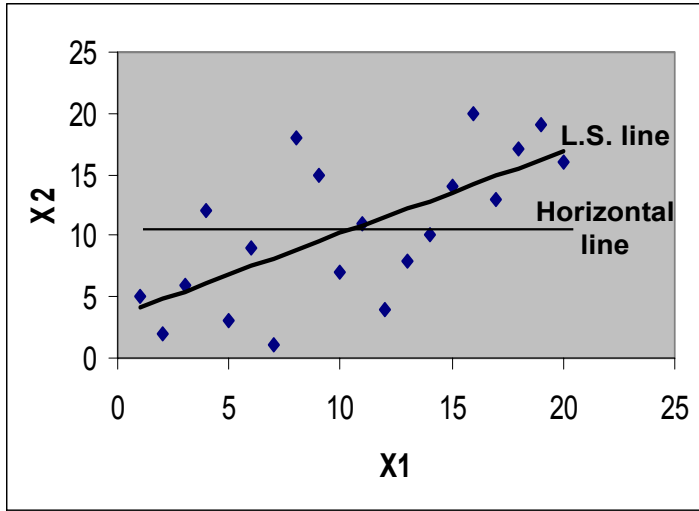


Figure 4: The least squares line along with the horizontal line.

The ratio between the deviations from the least squares fitted line and the deviations from the horizontal line is the measure used to express the correlation [4]. The stronger the correlation, the smaller the deviations from the least squares fitted line, thus the closer this ratio is to 0. Conversely, the weaker the correlation, the bigger the deviations from the least squares fitted line, thus the closer this ratio is to 1. Formally, this ratio is subtracted from 1 so a higher correlation will correspond to a higher value of this measure. Let us formalize this measure: Let (X_{1i}, X_{2i}) denote a point in the plane spanning the two variables; let \bar{X}_2 denote the mean of \bar{X}_2 (and the horizontal line as well); let \hat{X}_{2i} denote a predicted point of X_{2i} given X_{1i} , using the least squares fitted line equation: $\hat{X}_{2i} = a + bX_{1i}$. Thus the measure of the ratio is as brought in equation 3.

$$h = 1 - \frac{\sum_{i=1}^m (X_{2i} - \hat{X}_{2i})^2}{\sum_{i=1}^m (X_{2i} - \bar{X}_2)^2} \quad (3)$$

The denominator of the fraction (depicted in figure 5) is a familiar measure. Remember that the variance of \bar{X}_2 is: $V_2 = \sum_{i=1}^m (X_{2i} - \bar{X}_2)^2 / m$

Respectively, the numerator of the fraction measures the portion of variation (variance) of \bar{X}_2 , unexplained by the covariation with \bar{X}_1 (also depicted in figure 5). Thus, this ratio measures: 1-(independent_variation/total_variation), which is the amount of variation in \bar{X}_2 explained by the covariation with \bar{X}_1 .

What is the exact relationship between the correlation measured by h and the correlation given in equation 2? It turns out that $h = r_{X_1, X_2}^2$, termed the *coefficient of determination*. (See appendix A for detailed proof). Whereas r_{X_1, X_2} is measured using standard deviations units, h is measured using variance units. According to this relation between r_{X_1, X_2} and r_{X_1, X_2}^2 , while r_{X_1, X_2} can assume values between -1 to 1, r_{X_1, X_2}^2 can assume values between 0 to 1; while

$r_{X_1X_2}$ expresses the degree of linearity and its direction (positive or negative) between a pair of variables, $r_{X_1X_2}^2$ expresses the percentage of shared variation between these variables, regardless of the direction of the correlation. Going back to equation 3, an essential conclusion about $r_{X_1X_2}^2$ (and thus about correlation as well) can be drawn. A large value of $r_{X_1X_2}^2$ indicates strong co-variation, or better yet, significant mutual trend of variables in the data [5]. Since correlation is symmetric $r_{X_1X_2} = r_{X_2X_1} \Rightarrow r_{X_1X_2}^2 = r_{X_2X_1}^2$, thus h could have been measured using the vertical line: \bar{X}_1 , using the deviations relative to it: $(X_{1i} - \bar{X}_1)$, and the least squares fitted line could be expressed as: $\hat{X}_{1i} = a' + b'X_{2i}$ using the deviations relative to it: $(X_{1i} - \hat{X}_{1i})$. (For further details see appendix A).

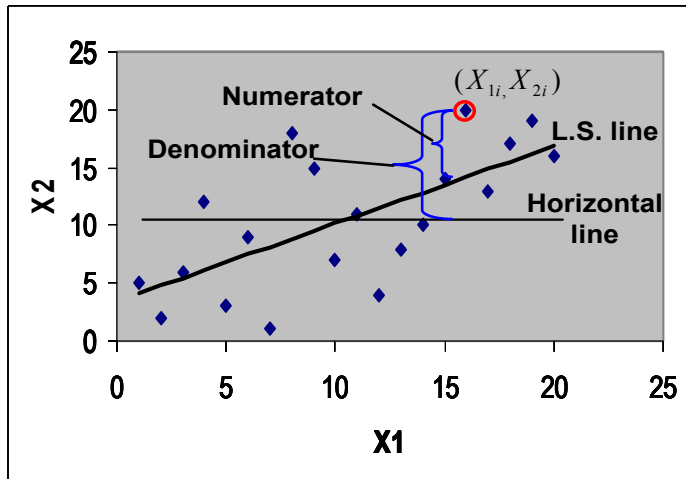


Figure 5: Deviations measurements from a point in the scatter plot.

Interestingly, the least squares fitted line actually maximizes the variance of the projections of the points upon it (figure 6). Therefore, correlation between variables corresponds to the degree of variance created by the projections of the points from which these variables have been measured. In relation to PCA, identifying 'shared' trends in the data is achieved by finding trend lines which directions maximize the variances of the projections of the data points upon them. Accomplishing this serves the goals of PCA of reducing redundancy and discovering meaningful patterns in the data.

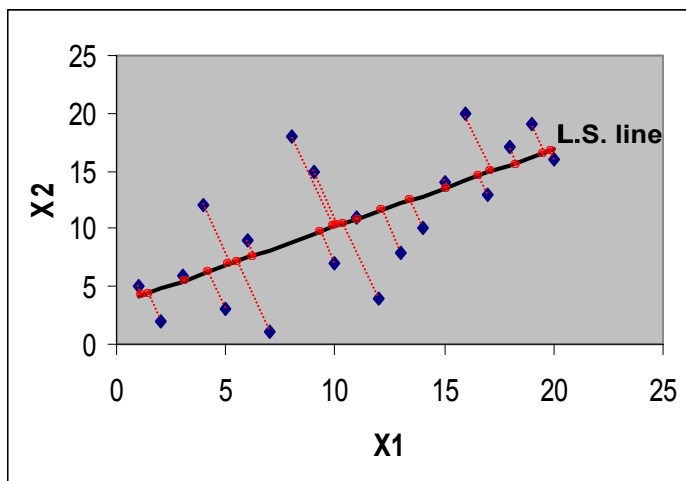


Figure 6: Projections of the point upon the least squares fitted line.

The vector approach for defining correlation

In this approach, each variable can be portrayed as a vector in R^m . Thus, the variation of the variables across our observations is indicated by the length and direction of the vector [4]. Given a pair of vectors, their relative magnitude and orientation towards each other shows how they vary together [4]. As used in physics, vectors pointed at a similar direction are like forces working together, which is to say they are varying together (figure 7a); vectors that are oppositely directed have inversed efforts, thus are working against each other, which is to say their variation is opposite (figure 7b); vectors that are at right angles, work independently, which is to say their variation is independent (figure 7c).

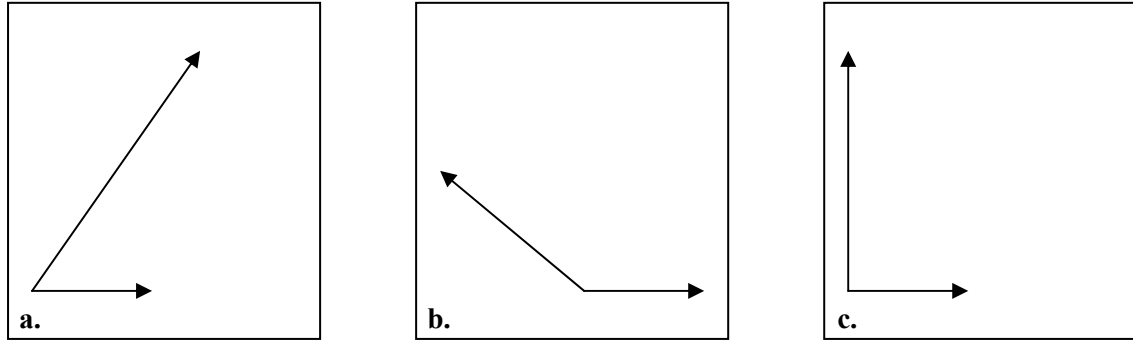


Figure 7: Relative orientations of a pair of vectors.

Although not a precondition, it is common to work with vectors of equal lengths. This could be achieved by *normalization*, which means dividing each coordinate of the vector by the vector's length. As a result the vector's length becomes 1 and each coordinate is calculated as presented in equation 4.

$$X_{li}^n = \frac{X_{li}}{\sqrt{\sum_{j=1}^m X_{lj}^2}}; \text{ For } i = 1, \dots, m \quad (4)$$

The superscript n denotes it is a normalized value.

In addition, we still like to remove the differences in the magnitudes of the coordinates between different vectors. Each vector is represented by a different coordinate system since it is measured by different units. By subtracting the mean of the vector's coordinates from each coordinate we simply translate the origin of the coordinate system, representing the vector, to the mean of the vector, which equals 0 (see equation 5). As a result, all the vectors are translated to the same origin of the same coordinate system. Actually, this procedure is analogous to standardizing the deviations used in the computational approach.

$$X_{ki}^t = \frac{(X_{ki}^n - \bar{X}_k^n)}{\sqrt{\sum_{i=1}^m (X_{ki}^n - \bar{X}_k^n)^2}} \text{ for every value } i \text{ of every vector } k \quad (5)$$

The superscript t denotes the vector is transformed.

Now we are simply left with dealing with the relative orientation between these transformed vectors. The essence of covariation between a pair of transformed vectors is their co-directionality, which means the angle between them. It is the *cosine* of this angle that expresses this relation in terms of correlation units. The cosine of the angle θ between two transformed vectors, P and Q , is computed as seen in equation 6.

$$P \cdot Q = \|P\| \|Q\| \cos(\theta_{PQ}) \quad (6)$$

Replacing P and Q with our transformed, normalized vectors yields the expression given in equation 7.

$$\text{Cos}(\theta_{X'_1 X'_2}) = \frac{\sum_{i=1}^m (X_{1i}^n - \bar{X}_1^n)(X_{2i}^n - \bar{X}_2^n)}{\sqrt{\left(\sum_{i=1}^m (X_{1i}^n - \bar{X}_1^n)^2\right)\left(\sum_{i=1}^m (X_{2i}^n - \bar{X}_2^n)^2\right)}} \quad (7)$$

What is the connection between the cosine measurement and the numerical measurement of correlation? This connection is developed in equation 8.

$$\text{Cos}(\theta_{X'_1 X'_2}) = \frac{\sum_{i=1}^m (X_{1i}^n - \bar{X}_1^n)(X_{2i}^n - \bar{X}_2^n)}{\sqrt{\left(\sum_{i=1}^m (X_{1i}^n - \bar{X}_1^n)^2\right)\left(\sum_{i=1}^m (X_{2i}^n - \bar{X}_2^n)^2\right)}} = \frac{\sum_{i=1}^m (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\left(\sum_{i=1}^m (X_{1i} - \bar{X}_1)^2\right)\left(\sum_{i=1}^m (X_{2i} - \bar{X}_2)^2\right)}} \quad (8)$$

$$\text{Thus, } r_{X_1 X_2} = \text{Cos}(\theta_{X'_1 X'_2})$$

A Theoretical Illustration of PCA

Consider a dataset where n parametric variables (x_1, x_2, \dots, x_n) were collected from m observations. The aim of PCA, hence, is to identify $k < n$ (usually $k = 2$ or 3) new variables (that will turn out to be the principal components) that determine a large portion of the information stored in the data, by accounting for the highest covariations possible in it.

For simplicity, we will consider $n = 3$ variables, collected from $m = 20$ observations, so we can visualize the process of representing the initial variables and the procedure of identifying and visualizing the principal components. Table 1 presents the dataset in what is termed the *score matrix* [6].

Table 1: Example dataset, 3 parametric variables obtained for 20 observations.

Observation	V ₁	V ₂	V ₃
a	100	8	5
b	228	21	2
c	341	31	10
d	472	40	15
e	578	48	3
f	699	60	12
g	807	71	14
h	929	79	16
i	1040	92	18
j	1160	101	38
k	1262	109	28
l	1376	121	32
m	1499	128	35
n	1620	143	28
o	1722	150	30
p	1833	159	15
q	1948	172	12
r	2077	181	33
s	2282	190	23
t	2999	202	29

Our dataset creates a cloud of 20 points in a 3D space. Each measured variable corresponds to one of the axes which represent the projections of the points. This cloud is presented by the scatter plot in figure 8.

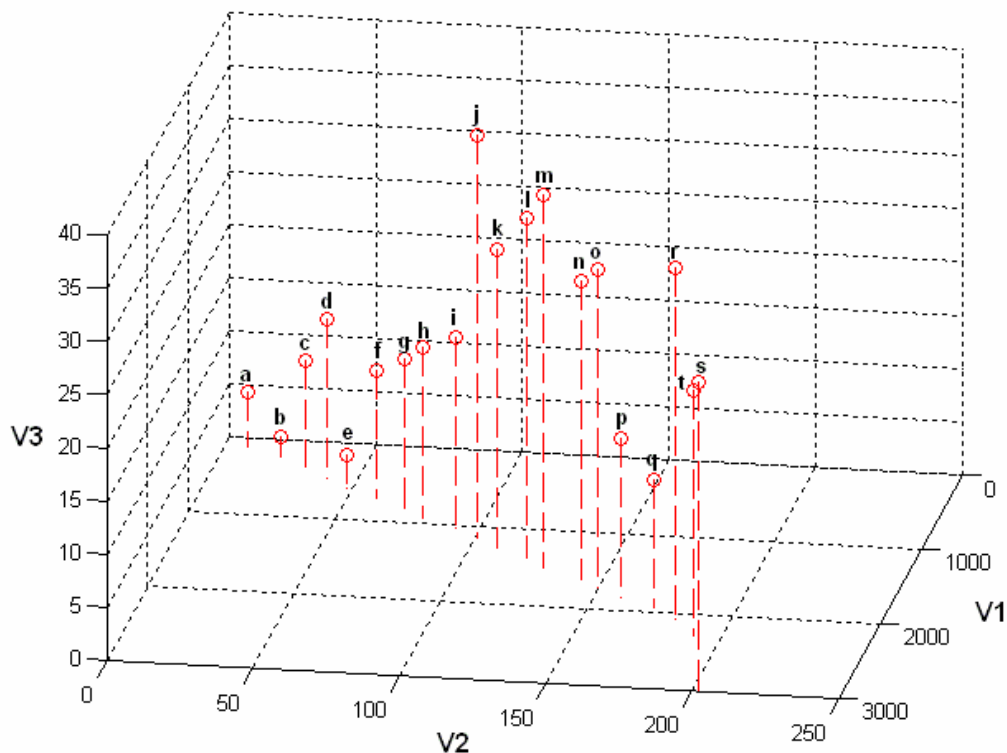


Figure 8: Scatter plot of the example dataset. The X-axis is defined by V₁, the Y-axis is defined by V₂, and the Z-axis is defined by V₃. Points are indicated with the corresponding observations letters. This example was cited from [7].

Considering the projections of the points on the plane defined by V1 and V2 (figure 9a), it is possible to conclude that V1 and V2 seem to be linearly related, and actually the line $V2 = V1$ approximately determines the direction by which V1 and V2 vary. However, when considering the entire space, it is not clear whether V3 is related either to V1 or to V2 as could be understood by looking at the projections of the points on the V1, V3 plane and the V2, V3 plane (figures 9b and 9c, respectively).

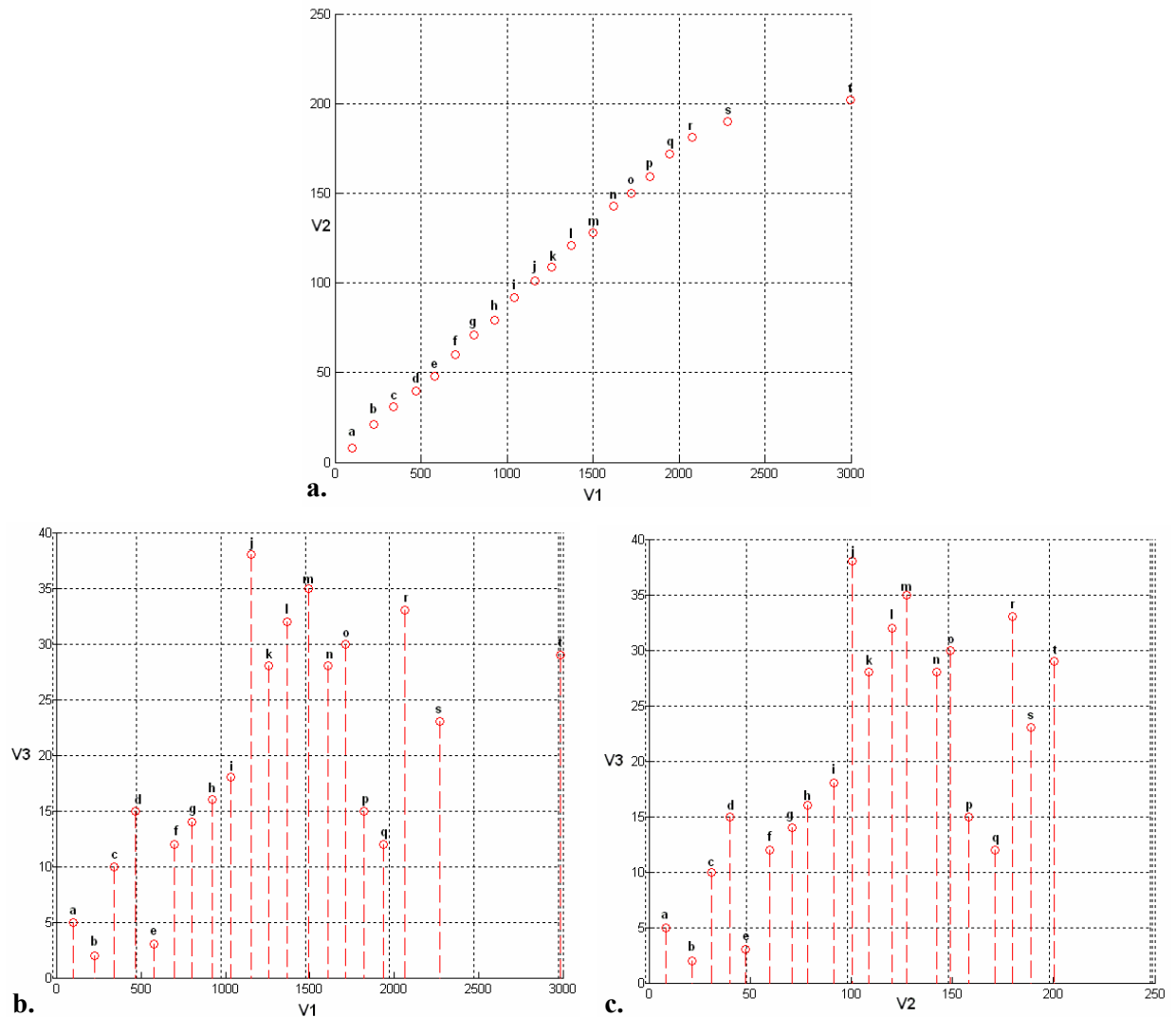


Figure 9; a. Projections of the points of the dataset on the V1,V2 plane; b. Projections of the points of the dataset on the V1,V3 plane; c. Projections of the points of the dataset on the V2,V3 plane.

Carrying on with the analysis, the score matrix should be standardized to a Z-score matrix (as explained above). As a result, our variables are transformed into numbers with standard deviation = 1 units and mean = 0. The relative location of the points remains the same, but the numbers are now comparable, as this procedure is a linear transformation performed on the dataset [2]. Table 2 presents the standardized dataset and figure 10 presents the 3D scatter plot defined by the standardized dataset.

Table 2: Standardized dataset.

Observation	V ₁	V ₂	V ₃
a	-1.51	-1.63	-1.34
b	-1.34	-1.41	-1.61
c	-1.19	-1.25	-0.89
d	-1.02	-1.10	-0.44
e	-0.88	-0.96	-1.52
f	-0.72	-0.76	-0.71
g	-0.58	-0.58	-0.53
h	-0.42	-0.44	-0.35
i	-0.27	-0.22	-0.17
j	-0.12	-0.07	1.63
k	0.02	0.06	0.73
l	0.17	0.26	1.09
m	0.33	0.38	1.36
n	0.49	0.63	0.73
o	0.62	0.75	0.91
p	0.77	0.90	-0.44
q	0.92	1.12	-0.71
r	1.09	1.27	1.18
s	1.36	1.42	0.28
t	2.30	1.62	0.82

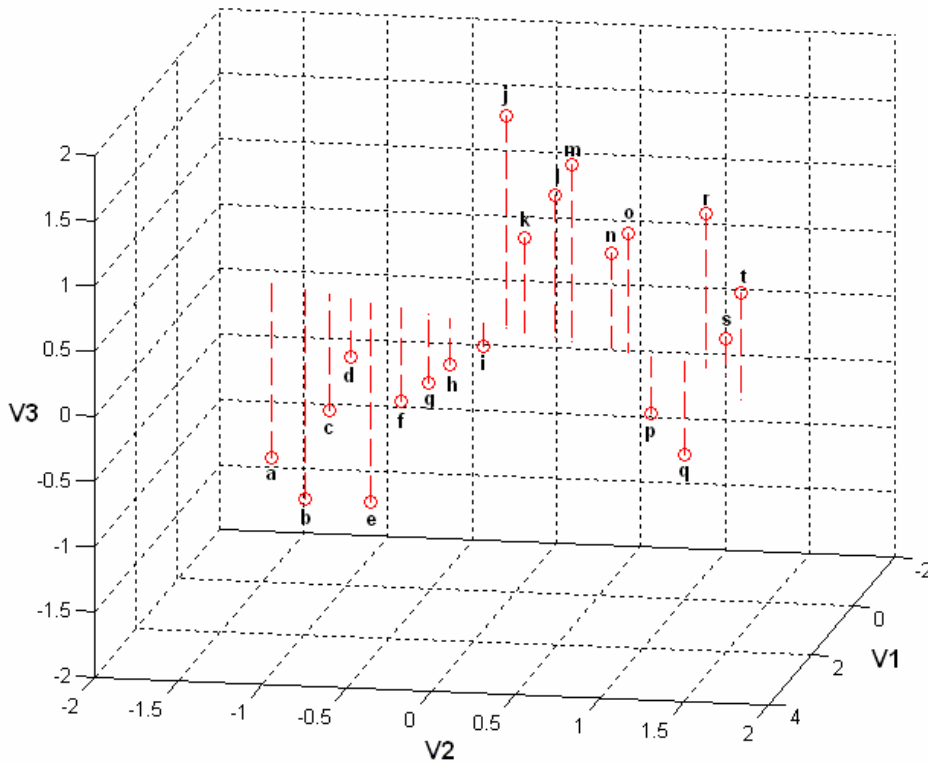


Figure 10: Scatter plot of the standardized dataset.

Now we are left with identifying one or two principal components. As each one of the measured variables corresponds to one of the axes representing the projections of the points, the principal components, to be identified, also correspond to axes of a new coordinate system we wish to find. The axis corresponding to the direction of the first principal component will be represented by a line that is required to minimize the sum of squared perpendicular distances from it to each point in the dataset. This least squares fitted line represents a line of correlation, thus the axis corresponding to the direction of the first principal component will

be aligned with the clearest trend in the data. Put another way, the least squares fitted line is the line maximizing the variance of the projections of the points upon it.

The direction of the second principal component (or any additional principal component) must also minimize the sum of squared perpendicular distances of the dataset points to it, under a single constraint: it must be linearly independent from the direction representing the first (or previous) principal component(s) identified. This means that each principal component accounts for an independent trend in the data, and hence they are uncorrelated. (The formal proof for this statement along with an explanation about the technique used to identify the principal components is brought in the Computational Steps section, step 3, below). An additional constraint requires the directions representing the principal components to pass through the origin (also called the centroid in this case, which is the mean of the standardized dataset [8]). This constraint stems from our desire to use these lines as the axes of the new coordinate system we wish to transform our data into. Since we have initially standardized our dataset, the directions of the principal components to be identified naturally pass through the centroid.

Based on the requirements given above, the directions of two principal components were identified according to figure 8, as seen in figure 11.

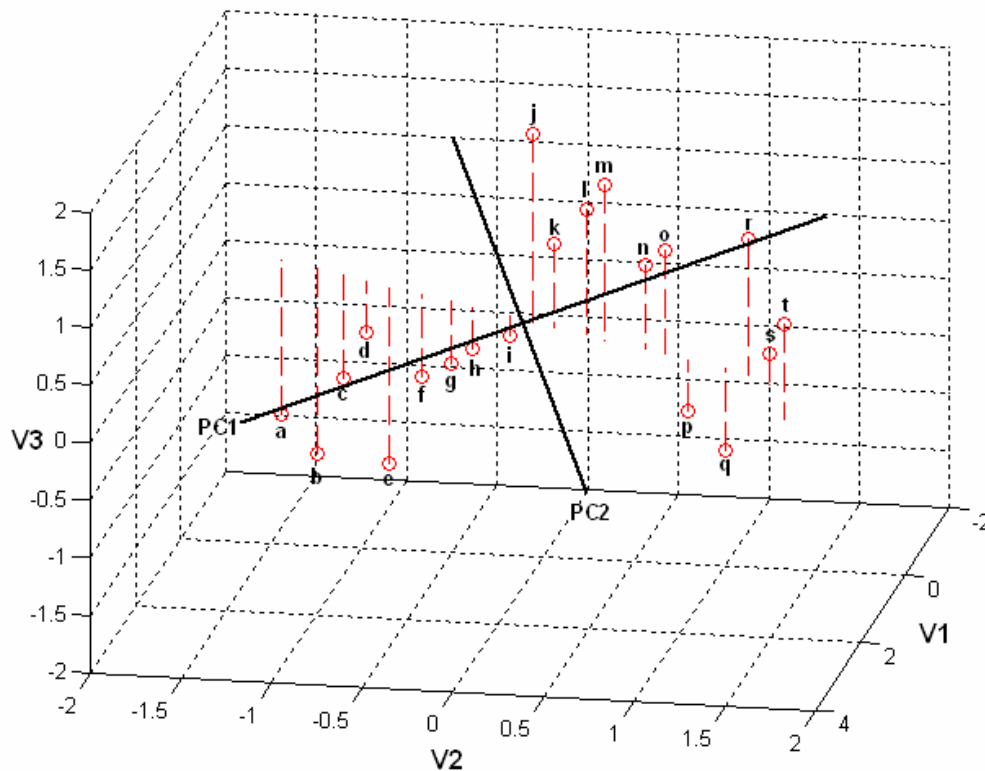


Figure 11: PC1 and PC2 estimation.

It is possible to see in figure 11 that:

1. The direction of the first principal component, PC1, passes through the centroid (0,0,0) and forms a least squares fitted line to what seems like the major trend in the data.
2. The direction of the second principal component, PC2, passes through the centroid as well, and is orthogonal to PC1, thus forms a least squares fitted line to a rather minor trend in the data, which is independent of the trend captured by PC1.

Rotating the plane spanned by PC1 and PC2 generates a 2D representation of the projections of the data points (figure 12). In this representation PC1 forms the *X*-axis (where the positive direction is that heading left in figure 10) and PC2 forms the *Y*-axis (where the positive direction is that heading down in figure 11).

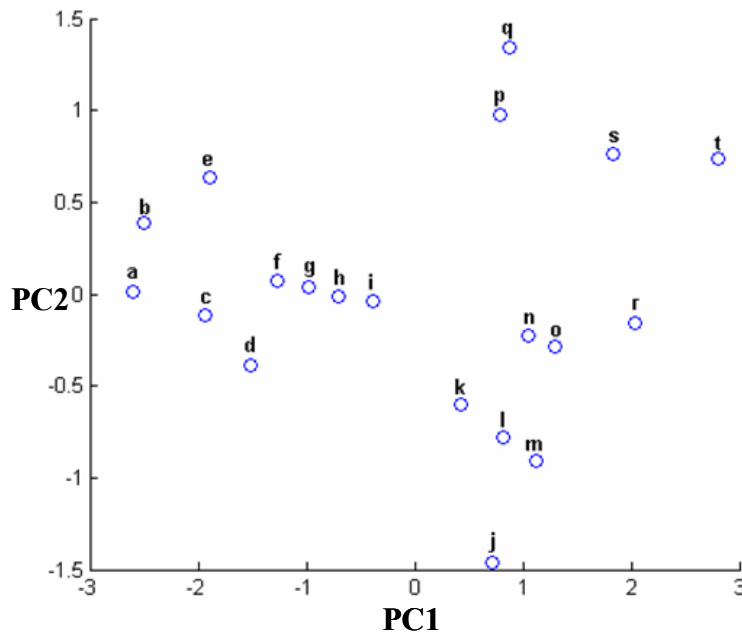


Figure 12: 2D plot generated by the PCA.

In conclusion, the dimensionality of the data has been reduced to 2 independent/uncorrelated variables without sacrificing much accuracy. The reduced dimension representation (figure 12, cited from [7]), is easier to interpret than the 3D plot presented in figure 8. For example, observations *n* and *o* seem to be related according to both PC1 and PC2, and observations *f*, *g*, *h* and *i* seem to be closely related according to both PC1 and PC2, as well. The complete interpretation of this resultant diagram is left for the user that needs to elucidate the underlying reasons responsible for the locations of the points.

It is worth noting that the relations between the points (figure 12) could have been identified directly from the 3D scatter plot, brought in figure 8, with the need of extra effort. However, this illustration is rather simplistic and PCA is intended for much more complicated cases with datasets consisting of multiple dimensions ($n \gg 3$).

Computational Steps

This section describes the computational steps that need to be accomplished in order to perform PCA, as presented superficially by the illustration given above.

The input to PCA is a dataset, held in a score matrix A , comprised of n variables (usually presented in columns) collected from m observations (usually presented in rows).

Step 1: Data standardization

As discussed above, this step is performed in order to avoid dealing with variables measured by different units. The mean and standard deviation are computed for each variable (column). The mean in each column, is reduced from every value in that column and the result is divided by the corresponding standard deviation. The standardized score matrix is annotated \bar{A} .

Step 2: Computing correlations between the variables

This step computes the correlation coefficient between each pair of variables in the dataset using equation 2. The motivation for executing this step is rather unclear at this stage, but a thorough explanation is provided in the next step.

As explained above, computing the correlation coefficient between each pair of unstandardized variables is identical to computing the covariance between each pair of standardized variables. Thus we can directly use equation 1.

Once the covariance measures have been computed for each pair of variables, they are stored in a matrix (termed the *covariance matrix*) that will serve the next steps. The dimensions of this matrix are $n \times n$, where cell (i,j) holds the value of $COV(\bar{X}_i, \bar{X}_j)$, where \bar{X}_i denotes variable i . Since covariance is symmetric as correlation is, $COV(\bar{X}_i, \bar{X}_j) = COV(\bar{X}_j, \bar{X}_i)$, hence the covariance matrix turns out to be symmetric.

Step 3: The connection between the data, the covariance matrix, its eigenvalues, its eigenvectors, the principal components, and computing all of them

As explained above, PCA searches for the directions of the principal components. These directions are vectors generate by performing a linear transformation on the data, originally spanned by n , correlated to a certain degree, variables, into n uncorrelated new variables. However, since PCA aims at reducing dimensionality, only $k < n$ of the new variables will be used eventually to present the data. The transformation used is linear combination on the n original variables, thus the objective of this step is to obtain the linear combinations which directions constitute correlation trend lines in the data [9] (equation 9).

$$\begin{aligned}
\bar{Y}_1 &= \bar{l}_1^T \cdot \bar{A} = l_{11}\bar{X}_1 + l_{12}\bar{X}_2 + \dots + l_{1n}\bar{X}_n \\
\bar{Y}_2 &= \bar{l}_2^T \cdot \bar{A} = l_{21}\bar{X}_1 + l_{22}\bar{X}_2 + \dots + l_{2n}\bar{X}_n \\
&\cdot \\
&\cdot \\
&\cdot \\
\bar{Y}_k &= \bar{l}_k^T \cdot \bar{A} = l_{k1}\bar{X}_1 + l_{k2}\bar{X}_2 + \dots + l_{kn}\bar{X}_n
\end{aligned} \tag{9}$$

Each coefficient vector \bar{l}_i^T generates a linear combination on \bar{A} which yields a new variable, \bar{Y}_i of dimension m , presented in the n dimension space. Each \bar{l}_i^T is required to be of magnitude = 1 in order to be one of the dimensions of the basis spanning \bar{A} . In addition, for each \bar{Y}_i to be considered as a principal component, the variance of the projections of the data points upon it need to be maximized. Altogether the following terms need to be satisfied:

1. The first principal component, \bar{Y}_1 is required to maximize $Var(\bar{l}_1^T A)$, subject to the constraint: $\|\bar{l}_1^T\| = 1$.
 2. The second principal component, \bar{Y}_2 is required to maximize $Var(\bar{l}_2^T A)$, subject to the constraints: $\|\bar{l}_2^T\| = 1$, and: $COV(\bar{l}_1^T A, \bar{l}_2^T A) = 0$.
- And in general, the i^{th} principal component, \bar{Y}_i is required to maximize $Var(\bar{l}_i^T A)$, subject to the constraints: $\|\bar{l}_i^T\| = 1$, and: $COV(\bar{l}_i^T A, \bar{l}_j^T A) = 0$, for all $i < j$.

The definition of $Var(\bar{Y}_i)$ is brought in equation 10 below.

$$\begin{aligned}
Var(\bar{Y}_i) &= Var(\bar{l}_i^T A) = E[(\bar{l}_i^T (A - E(A)))^2] \\
&= E[(\bar{l}_i^T (A - E(A))(A - E(A))^T \bar{l}_i)] \\
&= \bar{l}_i^T E[(A - E(A))(A - E(A))^T \bar{l}_i] = \bar{l}_i^T C \bar{l}_i
\end{aligned} \tag{10}$$

The term $E[(A - E(A))(A - E(A))^T]$ from equation 10 is actually the covariance matrix created in step 2, annotated here by C . The proper formulation of the problem of extracting up to k principal components therefore could be expressed as brought in equation 11 below.

$$\begin{aligned}
&\text{a. } \max \{ \bar{l}_i^T C \bar{l}_i \} \\
&\text{s.t.:} \\
&\|\bar{l}_i^T\| = \bar{l}_i^T \bar{l}_i = 1, \forall i. \\
&\text{b. } \max \{ \bar{l}_j^T C \bar{l}_j \} \\
&\text{s.t.:} \\
&1. \quad \|\bar{l}_j^T\| = \bar{l}_j^T \bar{l}_j = 1, \forall j. \\
&2. \quad \bar{l}_j^T \bar{l}_i = 0, \forall j \neq i.
\end{aligned} \tag{11}$$

The eigenvectors of C are scaled to have size = 1, they all pass through the origin, and they are all orthogonal to one another (or in short orthonormal). Thus, they satisfy all the terms presented in both optimization problems (equation 11, a and b). The technique for solving such optimization problems (linearly constrained) involves a construction of a LaGrangian function, using LaGrange multipliers [10, 11]. An informal formulation of this procedure is brought in equation 12 and in the proceeding paragraph.

$$\begin{aligned} \text{a. (for 11a): } L(\bar{l}_i, \lambda_i) &= \bar{l}_i^T C \bar{l}_i - \lambda_i (\bar{l}_i^T \bar{l}_i - 1); \text{ Find } \bar{l}_i \text{ and } \lambda_i \text{ that maximize } L. \\ \text{b. (for 11b): } L(\bar{l}_j, \lambda_j, \delta) &= \bar{l}_j^T C \bar{l}_j - \lambda_j (\bar{l}_j^T \bar{l}_j - 1) - \delta (\bar{l}_j^T \bar{l}_i); \end{aligned} \quad (12)$$

Find \bar{l}_j, λ_j and δ that maximize L .

Taking the partial derivative $\partial L / \partial \bar{l}_i$ for 12a obtains the condition: $\bar{l}_i^T C = \lambda_i \bar{l}_i^T$. By postmultiplying each side of this condition by \bar{l}_i we get: $\bar{l}_i^T C \bar{l}_i = \lambda_i$. Although not proven here, it is known from matrix algebra that the parameters: \bar{l}_i^T and λ_i that satisfy this condition are an eigenvalue and a corresponding eigenvector of C [12, 13]. Thus the optimum of the original objective function (equation 11a) is reached by assigning the maximal eigenvalue to λ_i and the corresponding eigenvector to \bar{l}_i .

Taking the partial derivative $\partial L / \partial \bar{l}_j^T$ for 12b obtains the condition: $C \bar{l}_j - \lambda_j \bar{l}_j - \delta \bar{l}_i = 0$. By premultiplying each side of this condition by \bar{l}_i^T we get: $\bar{l}_i^T C \bar{l}_j - \bar{l}_i^T \lambda_j \bar{l}_j - \delta \bar{l}_i^T \bar{l}_i = 0$. According to the definition of covariance: $\bar{l}_i^T C \bar{l}_j = \bar{l}_i^T \bar{l}_j = 0$, hence we get $\delta = 0$. As a result, the condition: $C \bar{l}_j = \lambda_j \bar{l}_j$ is obtained again, so once more the eigenvector-eigenvalue pair of C is reached, however $\lambda_j \neq \lambda_i$ and $\bar{l}_j \neq \bar{l}_i$. Using induction, it can be proven that up to k ($k < n$, where n is the rank of C) eigenvector-eigenvalue pairs can be found this way [9].

In conclusion, we have seen that the coefficients generating the linear combinations that transform the original variables into uncorrelated variables are the eigenvectors of the covariance matrix. The fraction of an eigenvalue out of the sum of all eigenvalues, of the covariance matrix, represents the amount of variation accounted by the corresponding eigenvector. The underlying intuition for this statement is that the magnitude of each eigenvalue represents the length of the corresponding eigenvector. The length of the eigenvector is a measure of the degree of common variation it accounts for in the data or better yet the intensity of a trend in the data. Empirically, when PCA works well, the first two eigenvalues usually account for more than 50% of the total variation in the data.

Trying to put it graphically, figure 13 below, presents the plot given in figure 11, where an ellipse, centered at (0,0,0), is constructed. PC1 and PC2 form the major and minor axes, respectively, such that this ellipse encloses the amount of variation accounted by both PCs.

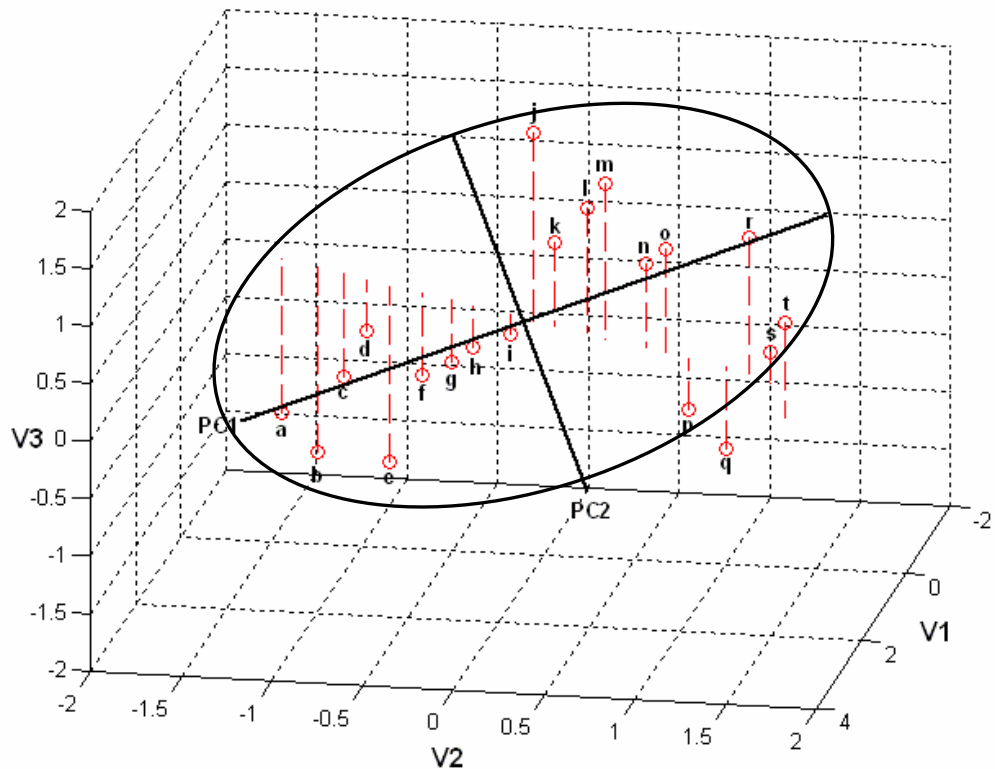


Figure 13: The scatter plot from figure 10, where the directions of the principal components PC1 and PC2, form the axes of the constructed ellipse.

The eigenvalues of the covariance matrix correspond to the length of the axes of the ellipse. The corresponding eigenvectors are coordinates, defining the direction of the axes. Stressing out the main idea of PCA again, the stronger the correlation in the data in a certain direction, the bigger the corresponding eigenvalue of the covariance matrix, the longer the axis is in that direction.

The decomposition of an $n \times n$ squared matrix into eigenvalues and eigenvectors is known as *eigen decomposition* [13]. This procedure is not straight forward, hence is usually left to be done using fancy computer software.

Step 4: Projecting the data on the reduced space spanned by the principal components

As seen in the diagram presented in figure 12, PCA eventually reduces the dimensions of the data according to the number of principal components that cover a sufficient amount of variation in it. In addition, the orientation of the data is rotated such that the directions of the principal components correspond to the axes of the coordinate system presenting the data. Computationally, this is achieved by premultiplying the transposed score matrix with a transformation matrix. This transformation matrix is the matrix that consists of the eigenvectors of the covariance matrix (i.e. \vec{l}_i^T), arranged in rows in descending order of the corresponding eigenvalues. The number of eigenvectors in the transformation matrix determines the number of dimensions the data will be transformed into. If k eigenvectors were

chosen to be included, we get a $k \times n$ transformation matrix. The score matrix is comprised of n variables of dimension m , hence it has size of $m \times n$. Transposing the score matrix yields an $n \times m$ matrix. Hence, the product of: $(k \times n)(n \times m)$ matrices yields a $k \times m$ matrix, where the rows represent the principal components and the columns represent the data points.

Concluding, this procedure is a transformation that allows us to obtain a linear projection of our data, originally in R^n , onto R^k , where $k < n$. Along with reducing the data dimensions, the data is also projected in a different orientation. Altogether, this transformation presents the data in a manner that stresses out the trends in it facilitating its interpretation.

References

- .1 http://149.170.199.144/multivar/pca_graf.htm.
- .2 <http://ordination.okstate.edu/glossary.htm#standardize>.
- .3 Bernstein, I.H., *Chapter 2: Some Basic Statistical Concepts*. Applied Multivariate Analysis: p. 22-46.
- .4 <http://www.mega.nu:8080/ampp/rummel/uc.htm>.
- .5 <http://www.psychstat.smsu.edu/introbook/sbk17.htm>.
- .6 Bernstein, I.H., *Chapter 6: Exploratory Factor Analysis*. Applied Multivariate Analysis: p. 157-182.
- .7 <http://ordination.okstate.edu/PCA.htm>.
- .8 <http://ordination.okstate.edu/glossary.htm#centroid>.
- .9 http://www.resample.com/xlminer/help/PCA/pca_intro.htm.
- .10 Shashua, A., *Intor. to Machine Learning. Lecture 9: Algebraic Representation I: PCA (scribe)*. 2003: p. 9-1 - 9-8.
- .11 Anderson, T.W., *Chapter 11: Principal Components*. An Itrouduction to Multivariate Statistical Analysis: p. 451-460.
- .12 <http://mathworld.wolfram.com/Eigenvalue.html>.
- .13 <http://mathworld.wolfram.com/Eigenvector.html>.
- .14 <http://mathworld.wolfram.com/CorrelationCoefficient.html>.
- .15 <http://www.efunda.com/math/leastquares/lstsqr1dcurve.cfm>.

Appendix A: Proving that: $h = r_{X_1X_2}^2$. (Cited from [14, 15]).

From equation 2 (in the text):

$$r_{X_1X_2} = \frac{\sum_{i=1}^m (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^m (X_{1i} - \bar{X}_1)^2 \sum_{i=1}^m (X_{2i} - \bar{X}_2)^2}} \quad (1)$$

$$\begin{aligned} \sum_{i=1}^m (X_{1i} - \bar{X}_1)^2 &= \sum_{i=1}^m X_{1i}^2 - 2\bar{X}_1 \sum_{i=1}^m X_{1i} + \sum_{i=1}^m \bar{X}_1^2 \\ &= \sum_{i=1}^m X_{1i}^2 - 2m\bar{X}_1^2 + m\bar{X}_1^2 = \sum_{i=1}^m X_{1i}^2 - m\bar{X}_1^2 = SS_{X_1X_1} \end{aligned} \quad (2)$$

Where $SS_{X_1X_1}$ denotes Sum of Squared values of each $X_{1i} \in X_1$ from \bar{X}_1 .

$$\begin{aligned} \sum_{i=1}^m (X_{2i} - \bar{X}_2)^2 &= \sum_{i=1}^m X_{2i}^2 - 2\bar{X}_2 \sum_{i=1}^m X_{2i} + \sum_{i=1}^m \bar{X}_2^2 \\ &= \sum_{i=1}^m X_{2i}^2 - 2m\bar{X}_2^2 + m\bar{X}_2^2 = \sum_{i=1}^m X_{2i}^2 - m\bar{X}_2^2 = SS_{X_2X_2} \end{aligned} \quad (3)$$

Where $SS_{X_2X_2}$ denotes Sum of Squared values of each $X_{2i} \in X_2$ from \bar{X}_2 .

$$\begin{aligned} \sum_{i=1}^m (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) &= \sum_{i=1}^m (X_{1i}X_{2i} - mX_{1i}\bar{X}_2 - m\bar{X}_1X_{2i} + \bar{X}_1\bar{X}_2) \\ &= \sum_{i=1}^m X_{1i}X_{2i} - m\bar{X}_1\bar{X}_2 = SS_{X_1X_2} \end{aligned} \quad (4)$$

Where $SS_{X_1X_2}$ denotes Sum of Squared values of each $(X_{1i}, X_{2i}) \in X_1, X_2$.

Putting 2,3 and 4 into 1 we get:

$$r_{X_1X_2} = \frac{SS_{X_1X_2}}{\sqrt{SS_{X_1X_1}SS_{X_2X_2}}} \quad (5)$$

Let us develop the least squares line equations:

1. For $\hat{X}_{2i} = a + bX_{1i}$ find a and b that define the line with the least square error:

$$\sum_{i=1}^m (X_{2i} - \hat{X}_{2i})^2 = \sum_{i=1}^m (X_{2i} - a - bX_{1i})^2 = \min \quad (6)$$

Deriving 6 by a and b we get:

$$\frac{\partial[\sum_{i=1}^m (X_{2i} - a - bX_{1i})^2]}{\partial a} = -2\sum_{i=1}^m [X_{2i} - (a + bX_{1i})] = 0 \quad (7)$$

$$\frac{\partial[\sum_{i=1}^m (X_{2i} - a - bX_{1i})^2]}{\partial b} = -2\sum_{i=1}^m X_{1i}[X_{2i} - (a + bX_{1i})] = 0$$

From 7 we get:

$$\begin{aligned} \sum_{i=1}^m X_{2i} &= a\sum_{i=1}^m 1 + b\sum_{i=1}^m X_{1i} \\ \sum_{i=1}^m X_{2i}X_{1i} &= a\sum_{i=1}^m X_{1i} + b\sum_{i=1}^m X_{1i}^2 \end{aligned} \quad (8)$$

Solving 8 (2 equations with 2 parameters) we get:

$$\begin{aligned} a &= \frac{(\sum_{i=1}^m X_{2i})(\sum_{i=1}^m X_{1i}^2) - (\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{1i}X_{2i})}{m\sum_{i=1}^m X_{1i}^2 - (\sum_{i=1}^m X_{1i})^2} \\ b &= \frac{m\sum_{i=1}^m X_{1i}X_{2i} - (\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{2i})}{m\sum_{i=1}^m X_{1i}^2 - (\sum_{i=1}^m X_{1i})^2} \end{aligned} \quad (9)$$

2. For $\hat{X}_{1i} = a' + b'X_{2i}$ find a' and b' that define the line with the least square error:

$$\sum_{i=1}^m (X_{1i} - \hat{X}_{1i})^2 = \sum_{i=1}^m (X_{1i} - a' - b'X_{2i})^2 = \min \quad (10)$$

Thus, deriving 10 by a' and b' :

$$\frac{\partial[\sum_{i=1}^m (X_{1i} - a' - b'X_{2i})^2]}{\partial a'} = -2\sum_{i=1}^m [X_{1i} - (a' + b'X_{2i})] = 0 \quad (11)$$

$$\frac{\partial[\sum_{i=1}^m (X_{1i} - a' - b'X_{2i})^2]}{\partial b'} = -2\sum_{i=1}^m X_{2i}[X_{1i} - (a' + b'X_{2i})] = 0$$

From 11 we get:

$$\begin{aligned}\sum_{i=1}^m X_{1i} &= a' \sum_{i=1}^m 1 + b' \sum_{i=1}^m X_{2i} \\ \sum_{i=1}^m X_{1i} X_{2i} &= a' \sum_{i=1}^m X_{2i} + b' \sum_{i=1}^m X_{2i}^2\end{aligned}\tag{12}$$

Solving 12 (2 equations with 2 parameters) we get:

$$\begin{aligned}a' &= \frac{(\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{2i}^2) - (\sum_{i=1}^m X_{2i})(\sum_{i=1}^m X_{2i} X_{1i})}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{2i})^2} \\ b' &= \frac{m \sum_{i=1}^m X_{2i} X_{1i} - (\sum_{i=1}^m X_{2i})(\sum_{i=1}^m X_{1i})}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2}\end{aligned}\tag{13}$$

Let us further develop the expressions for b and b' , from 9 and 13, respectively:

$$\begin{aligned}b &= \frac{m \sum_{i=1}^m X_{1i} X_{2i} - (\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{2i})}{m \sum_{i=1}^m X_{1i}^2 - (\sum_{i=1}^m X_{1i})^2} = \frac{m \sum_{i=1}^m X_{1i} X_{2i} - m \bar{X}_1 m \bar{X}_2}{m \sum_{i=1}^m X_{1i}^2 - (m \bar{X}_1)^2} \\ &= \frac{\sum_{i=1}^m X_{1i} X_{2i} - m \bar{X}_1 \bar{X}_2}{\sum_{i=1}^m X_{1i}^2 - m \bar{X}_1^2} = \frac{SS_{X_1 X_2}}{SS_{X_1 X_1}}\end{aligned}\tag{14}$$

$$\begin{aligned}b' &= \frac{m \sum_{i=1}^m X_{2i} X_{1i} - (\sum_{i=1}^m X_{2i})(\sum_{i=1}^m X_{1i})}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2} = \frac{m \sum_{i=1}^m X_{1i} X_{2i} - m \bar{X}_1 m \bar{X}_2}{m \sum_{i=1}^m X_{2i}^2 - (m \bar{X}_2)^2} \\ &= \frac{\sum_{i=1}^m X_{1i} X_{2i} - m \bar{X}_1 \bar{X}_2}{\sum_{i=1}^m X_{2i}^2 - m \bar{X}_2^2} = \frac{SS_{X_1 X_2}}{SS_{X_2 X_2}}\end{aligned}\tag{15}$$

Thus far it can be concluded that:

$$r_{X_1 X_2} = \sqrt{bb'}\tag{16}$$

From equation 3 (in the text):

$$h = 1 - \frac{\sum_{i=1}^m (X_{2i} - \hat{X}_{2i})^2}{\sum_{i=1}^m (X_{2i} - \bar{X}_2)^2} \quad (17)$$

Let us denote the numerator from 17 as *SSE* (sum of squared errors). Thus:

$$SSE = \sum_{i=1}^m (X_{2i} - \hat{X}_{2i})^2 = \sum_{i=1}^m (X_{2i} - a - bX_{1i})^2 \quad (18)$$

From 9:

$$a = \frac{(\sum_{i=1}^m X_{2i})(\sum_{i=1}^m X_{2i}^2) - (\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{1i}X_{2i})}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2} \quad (19)$$

But:

$$\begin{aligned} \bar{X}_2 - b\bar{X}_1 &= \bar{X}_2 - \frac{m \sum_{i=1}^m X_{1i}X_{2i} - (\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{2i})}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2} \bar{X}_1 \\ &= \frac{m\bar{X}_2 \sum_{i=1}^m X_{2i}^2 - \bar{X}_2 (\sum_{i=1}^m X_{1i})^2 - m\bar{X}_1 \sum_{i=1}^m X_{1i}X_{2i} + \bar{X}_1 (\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{2i})}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2} \\ &= \frac{m\bar{X}_2 \sum_{i=1}^m X_{2i}^2 - m\bar{X}_1 \sum_{i=1}^m X_{1i}X_{2i} - \bar{X}_2 (\sum_{i=1}^m X_{1i})^2 + \frac{\sum_{i=1}^m X_{1i}}{m} (\sum_{i=1}^m X_{1i}) m\bar{X}_2}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2} \\ &= \frac{m\bar{X}_2 \sum_{i=1}^m X_{2i}^2 - m\bar{X}_1 \sum_{i=1}^m X_{1i}X_{2i}}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2} = \frac{(\sum_{i=1}^m X_{2i})(\sum_{i=1}^m X_{2i}^2) - (\sum_{i=1}^m X_{1i})(\sum_{i=1}^m X_{1i}X_{2i})}{m \sum_{i=1}^m X_{2i}^2 - (\sum_{i=1}^m X_{1i})^2} = a \end{aligned} \quad (20)$$

Hence:

$$\begin{aligned}
SSE &= \sum_{i=1}^m [X_{2i} - (\bar{X}_2 - b\bar{X}_1) - bX_{1i}]^2 = \sum_{i=1}^m [X_{2i} - \bar{X}_2 - b(X_{1i} - \bar{X}_1)]^2 \\
&= \sum_{i=1}^m (X_{2i} - \bar{X}_2)^2 - 2b \sum_{i=1}^m (X_{2i} - \bar{X}_2)(X_{1i} - \bar{X}_1) + b^2 \sum_{i=1}^m (X_{1i} - \bar{X}_1)^2 \\
&= SS_{X_2X_2} - 2bSS_{X_1X_2} + b^2SS_{X_1X_1}
\end{aligned} \tag{21}$$

Replacing the expression for b from 14, in 21, we get:

$$\begin{aligned}
SSE &= SS_{X_2X_2} - 2 \frac{SS_{X_1X_2}^2}{SS_{X_1X_1}} + \frac{SS_{X_1X_2}^2}{SS_{X_1X_1}^2} SS_{X_1X_1} = SS_{X_2X_2} - \frac{SS_{X_1X_2}^2}{SS_{X_1X_1}} \\
&= SS_{X_2X_2} \left(1 - \frac{SS_{X_1X_2}^2}{SS_{X_1X_1}SS_{X_2X_2}}\right) = SS_{X_2X_2} (1 - r_{X_1X_2}^2)
\end{aligned} \tag{22}$$

From 3 we know that: $\sum_{i=1}^m (X_{2i} - \bar{X}_2)^2 = SS_{X_2X_2}$. Thus:

$$\begin{aligned}
h &= 1 - \frac{\sum_{i=1}^m (X_{2i} - \hat{X}_{2i})^2}{\sum_{i=1}^m (X_{2i} - \bar{X}_2)^2} = 1 - \frac{SSE}{SS_{X_2X_2}} = 1 - \frac{SS_{X_2X_2} (1 - r_{X_1X_2}^2)}{SS_{X_2X_2}} \\
&= 1 - (1 - r_{X_1X_2}^2) = r_{X_1X_2}^2
\end{aligned} \tag{23}$$

THE END