THE UNIVERSITY OF HERTFORDSHIRE

BUSINESS SCHOOL

WORKING PAPER SERIES

University of Hertfordshire Business School
College Lane
Hatfield
Hertfordshire
AL10 9AB
United Kingdom

Estimating Invariant Principal Components Using Diagonal Regression

Michael Leznik, Chris Tofallis

Department of Management Systems
The Business School
University of Hertfordshire

We would like to express our gratitude to the following people:

- Special thanks to *Emeritus* Prof. Ian Jolliffe of Aberdeen University for reviewing this work.
- Prof. Olav Bjerkholt of Oslo University for helping with the bibliography and providing important historical information.
- Chris Chatfield, Reader in Statistics at the University of Bath, for his most valuable help and comments.
- Dr. Mariano Matilla García of Madrid University for the help and inspiration he has provided.
- Dr. Paul Taylor and Dr. Neil Spencer of Hertfordshire University for reviewing this work and making valuable comments.
- Ms. Julia Brodskaya for beautiful illustrations she has made for this work.

**The views expressed in this work are those of the authors and any errors are their responsibility*

Keywords: Principal Component Analysis (PCA), Invariant Principal Component Analysis, orthogonal regression, diagonal regression, invariant regression, geometric mean functional relationship, generalized eigenproblem.

Abstract

In this work we apply the method of diagonal regression to derive an alternative version of Principal Component Analysis (PCA). “Diagonal regression” was introduced by Ragnar Frisch (the first economics Nobel laureate) in his paper “Correlation and Scatter in Statistical Variables” (1928). The benefits of using diagonal regression in PCA are that it provides components that are scale-invariant (i.e. changing the units of measurement leads to an equivalent result), and which reflect both the correlation structure of the data set, and the variance structure as well. By contrast PCA based on the correlation matrix will only reflect the correlation structure of the data. The problem is formulated as a generalized eigen-analysis and is demonstrated using a numerical example which highlights some desirable properties of what we call Invariant Principal Components Analysis (IPCA).

Introduction

Principal Component Analysis (PCA) is quite widely used in different areas such as data compression, image processing, visualisation, exploratory data analysis, pattern recognition etc. One may find a chapter on PCA in numerous texts on multivariate analysis e.g. Rao (1952), Kendall (1965), Gnanadesikan (1977), Chatfield and Collins (1986). For a more detailed explanation there are books entirely dedicated to Principal Component Analysis: Dunteman (1989), Jolliffe (2002), Jackson (2003). PCA originated in some work by Karl Pearson (1901) around the turn of the 20th century. Frisch (1928) introduced his view on how to transform a set of statistical variables to an uncorrelated set. Hotelling (1933) developed the approach to Principal Component Analysis which prevails in most textbooks today. According to Jolliffe (2002) the central idea of Principal Component Analysis is to reduce the dimensionality of a data set which may consist of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. Thus principal component analysis is concerned with reducing the number of variables under consideration by using only the first few principal components and discarding those, which have small variances. When one is dealing with high dimensional data it is often necessary to reduce its dimensionality, either to reduce storage requirements, for speedier transmission of information, or to make further analysis easier. Typically, the computation time in statistical analysis grows at a rate which is exponentially related to the dimension of the data.

Kendall (1965) summarises the underlying idea: “A linear or orthogonal transformation is applied to the p variates $x_1, x_2, ..., x_p$ to produce a new set of uncorrelated variates $Z_1, Z_2, ..., Z_n$”. In general PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. The success of PCA is due to the following important properties:

- ✓ Principal components sequentially capture the maximum variability among the data, thus guaranteeing minimal information loss when lesser components are discarded.
- ✓ Principal components are uncorrelated, so one can talk about each of the Principal components without referring to the others, each one makes an independent contribution to accounting for the variance of the original variables Dunteman (1989).

ten Berge and Kiers (1996) consider the following three traditional approaches to PCA:

1. Components which themselves possess maximum variance.
2. Components which explain the maximum amount of variance in the original variables, by optimal least squares fitting.
3. Components which provide an optimal least squares fit of the covariance or correlation matrix of the original variables.

Jackson (2003) observes that as an alternative to traditional PCA one can compute one's own principal components based on subjective criteria, although one should investigate the properties of these components, particularly with regard to the extent to which they are correlated and the extent to which they account for variability in the original variables. Korhonen (1984) calculates, what he calls, Subjective Principal Components by maximizing the absolute values of the correlations between principal components and the variables important to one, and not by maximizing their variances. Another approach was proposed by Kaiser (1967): to obtain components by looking for linear combinations of the original variables of the form Z=XA, where A is chosen such that the trace, Tr(A), is maximized so that each column of X (i.e. one of the *p* original variables) is paired with a column of Z, the sum of correlations over all *p* pairs being as large as possible.

Devlin et al. (1981) distinguish two general approaches for principal components. The first is to view the problem as Pearson (1901) did, as one of fitting a sequence of mutually orthogonal hyperplanes, and to replace the criterion minimizing the sum of squares of perpendicular deviations of the observations from the fitted plane by other criteria possessing desirable properties. The second approach is to perform standard eigenanalysis computations on the different measures of multivariate dispersion. Work in this direction has been done by Campbell (1980), Devlin *et al.* (1981) and Mathews (1984), they were interested in robust measures i.e., robust covariance and correlation matrices. Different estimators were used to create a form of PCA which is not overly affected by atypical observations. The problem these authors wanted to solve was the scale dependency of principal components.

PCA is well known to be scale dependent and so some form of normalisation is required. The usual approach is to standardize variables so that they have zero mean and unit variance. The idea is that all the variables have equal importance, where importance is assumed to be measured by the variance. Jackson (2003) states that the choice of scale will determine the dispersion matrix used to obtain components. If no scaling is employed, the resultant matrix will be a second moment matrix; if the mean is subtracted it will be a covariance matrix; if data is in standardized units it will be a correlation matrix. The problem associated with the covariance matrix is that if one of the variables has greater variance than the others, then the first principal component will be more influenced by this variable. In an effort to make all variables equally 'important', the correlation matrix is used instead of the covariance matrix. Under this standardisation the resulting principal components will not be equivalent to those using covariances. More generally, changing the type of normalisation used will affect the resulting components. For example, dividing each variable by its mean, or by its inter-quartile range, or using logs, will all make the data dimensionless, but in each case the set of principal components obtained will not be equivalent to that from other normalisations. From this one may conclude that the PCA method is not unit invariant: changes of scale affect the components one ends up with. Kendall (1965) expressed this in geometrical language: "lines of closest fit found by minimizing the sum of squares of perpendicular distances are not invariant under change of scale". This difficulty has been well known since the introduction of PCA, and different methods have been suggested for dealing with it. Hotelling (1933) notices that: "since the set of variables is capable of transformations such as changes of units and other linear transformations, the ellipsoids may be stretched and squeezed in any way. The method of principal components can therefore be applied only if for each variable there exists a unit of measure of unique importance". Sokal and Rohlf (1981) point out that determining the slope of the major axis (principal axes) can be done if both variables are in the same units of measurement, but when two variables have different units of measurements the slope of the major axis is meaningless and another technique should be employed. When the correlation matrix is used, variables are standardized to have zero mean and unit standard deviation. Nevertheless, standardization does not solve the scale dependency problem, but just avoids it. It merely forces upon the user a unit of measurement equal to one standard deviation.

As an illustration, Loretan (1997) applies PCA in order to generate market risk scenarios. Significant correlation between different financial variables allows the use of PCA to reduce the dimensionality of the data. He notices that: "Since PCA is sensitive to the units of measurement of the data, we report our results both for the "raw" and for "standardized" (zero mean, unit variance) series. Standardization is found to have little qualitative effect except when groups of series with differing group variances, such as exchange rates and interest rates, are analysed". He notes that when a combination of stock market indices, exchange rates and long term interest rates are analysed for unstandardized series, the first PC explains 50% of the variance. However, upon standardization the influence of first PC is diminished to 26%. The explanation for this can be found in Jackson (2003) who says that "there is no one-to-one correspondence between the PCs obtained from a correlation matrix and those obtained from a covariance matrix". Gnanadesikan (1977) also states that "principal components of the covariance matrix are not the same as those of the correlation matrix, or of some other scaling according to measure of 'importance'". Another example of an application of PCA to finance can be found in Lardic et al. (2002) where PCA is used to analyse the interest rate exposure of fixed-income portfolios. The authors state that, depending on the choice of original variables (scaled or not scaled), different sensitivities (components) are obtained. This had the extremely important effect that the structure of the investment portfolio differed and hence the performance would be affected by the scales of the variables.

Gnanadesikan (1977) observes that for reasons of a statistical nature such as: interpretation, formal statistical inference and distribution theory, it is often preferable to work with PCA based on the covariance matrix. Healy (2000, p96) makes the point very strongly that "the common choice of the [correlation matrix] for analysis has little or no theoretical motivation"

Interpretational problems of PCA based on the correlation matrix are well described by Walker (1967) where she criticises Ahamad (1967). The data analysed by Ahamad (1967) consist of the number of offences, classified according to 18 categories. The first principal component Z_1, by definition, is a weighted sum of the number of crimes in the 18 categories. This is expressed as:

$$Z_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ...a_{1n}x_n$$

Where x_n is the number of crimes in category n and $\boldsymbol{a}$ is the eigenvector of the first PC. Ahamad (1967) performs PCA on the correlation matrix using the data and suggests that the first component Z_1 can be described as a crime rate. However, if every variable is standardized, one finds that, for example, the crime of larceny, with values of order 300,000 per year contributes to the weighted sum Z_1 about the same amount as robbery, with only about 1,000 crimes per year. Therefore, it is difficult to see what Z_1 measures. One can find it quite awkward to make equally important such different types of crimes as larceny and robbery. As result Walker (1967) suggests not to analyse this data using PCA, but to use a different technique.

Another problem is noted by Chatfield and Collins (1986): “Analysing the correlation matrix also makes it more difficult to compare results from two or more different samples”. The problem is that PCA based on the correlation matrix takes into account only the correlation structure of the data without paying any attention to the differences in variances. Suppose two different samples have an apparently similar correlation structure, but actually have quite different properties in terms of their variances. To compare such samples by looking at the correlations alone is not enough. What one would probably like to see are differences in coefficients (elements of eigenvectors) which reflect the variance differences in the data.

Work on weighted principal components has been presented by Meredith and Millsap (1985). They describe it as “an alternative approach to component analysis which lends itself to a broad characterization of equivalent classes of component solutions under metric transformation”. They notice that since the choice of scale for many psychological measurements is arbitrary, the scale-invariance properties of component solutions are of particular concern to psychologists. Meredith and Millsap (1985) introduce two different criteria generalized to allow weighting, the choice of weights determining the scale invariance properties of the resulting solution. However, as the authors point out in their work, two criteria are developed and are shown to lead to different component solutions. This fact suggests that both solutions are not unique to the data characteristics.

The underlying idea we are introducing in this paper is based on the method of diagonal regression introduced by Frisch (1928). This regression line possesses properties which we feel make it worthy of application to multivariate analysis. The problem will be formulated as a generalized eigenanalysis, where the identity matrix *I* will be replaced by a diagonal matrix *D* containing products of moments.

Diagonal[1] regression and invariant estimators

Since the nineteenth century different methods have been developed to fit a straight line when both variables are subject to error. The earliest work appears to be that due to Adcock (1877), who suggested minimizing the sums of squares of the normals i.e. *perpendicular distances* (orthogonal regression), from the data points to the line. Later, Pearson (1901) introduced and explained the same approach. Pearson advocated this approach in the knowledge that in many cases in physics and biology the "independent" variable is subject to just as much deviation or error as the "dependent" variable. According to Reed (1921): "in practically all cases of observed data, x is as subject to variation as y and it therefore, appears *a priori* that a better fitting straight line would be obtained if we define the word residual as the normal deviation of an observed point from the line. This definition assumes that an observed point fails to fall on the line due to an error in both x and y". It's worth noting that Pearson (1901) not only proposed such a "best fit line", but also observed that it passes through the direction of maximum data variation and coincides with the direction of the maximum (principal) axis of the correlation ellipsoid and perpendicular to the least (minor) axis of the correlation ellipsoid.

Wald (1940) notices that many objections can be raised against this method. First, there is no justification for minimizing the sums of squares of the normal deviates – why not in some other direction? Second, the straight line obtained by that method is not invariant under transformation of the coordinate system. A criticism against orthogonal regression can also be found in Frisch (1934), who states that if variates are not normalized, orthogonal regression is not even invariant to a change in units of measurement. Roos (1937) emphasizes the same point, summarizes different methods, and then proposes a general formula for fitting lines (and planes in case of more than two variables), which do not depend on the choice of the coordinate system. Jones (1937) gives a geometrical interpretation of Roos's general solution and some of the special cases. He arrives at the conclusion that the "true" relation between two variables would be:

[1] In "Correlation and Scatter in Statistical variables" Ragnar Frisch introduced two invariant regressions: "diagonal" and "composite". However, in "Statistical Confluence Analysis by Means of Complete Regression Systems" he does not make this distinction and unites both lines under the names "diagonal" or "true" regression. Later on Cobb (1939) and Samuelson (1942) refer to Frisch's invariant regression as "diagonal regression".

$$y = \frac{\sigma_y}{\sigma_x} x$$

where σ_y and σ_x are population standard deviations of variables *x* and *y* respectively and states that "Geometrically this regression line is the diagonal of the rectangle circumscribing the correlation ellipse" (Figure 1).

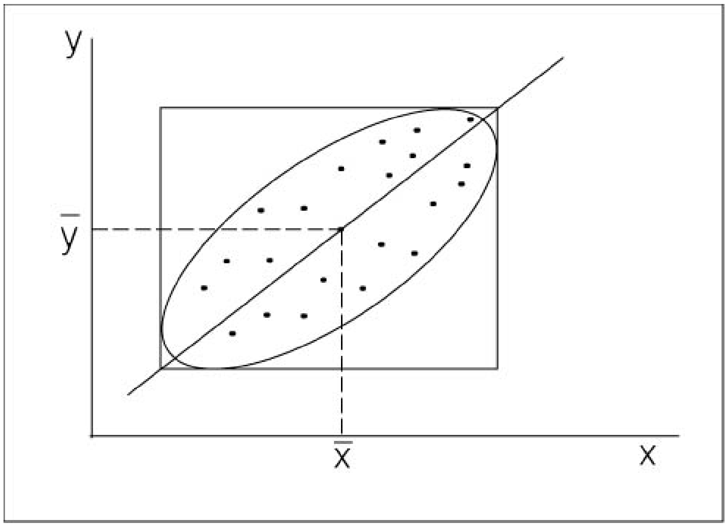

FIGURE 1

Woolley (1941) tackles the same problem again, but from a geometrical point of view. He presented a method of determining a straight-line regression by minimizing the summed absolute values of the areas of right-angled triangles formed by the data points and the regression line (Figure 2).

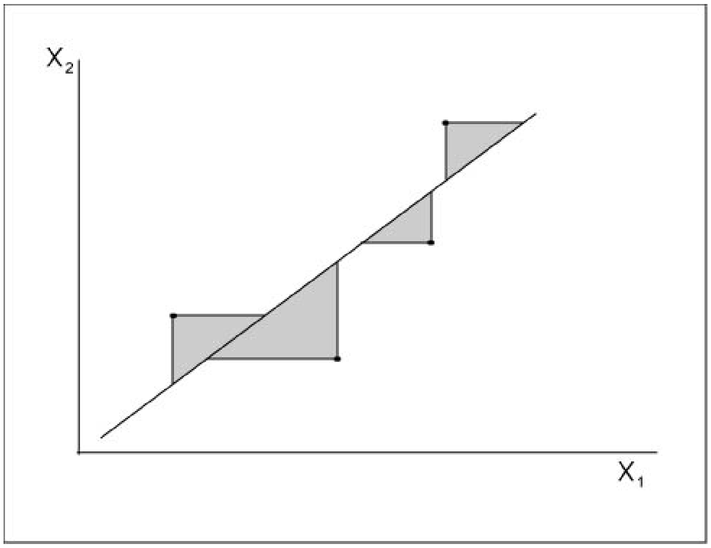

FIGURE 2

In this work, he proves that the slope of this "least triangles" regression in the case of a linear relationship of *y* on *x* is:

$$\pm\frac{\sigma_y}{\sigma_x},$$

and in the case of a relationship of *x* on *y:*

$$\pm\frac{\sigma_x}{\sigma_y}$$

The sign of the coefficients in both cases is determined by the sign of the correlation coefficient. Samuelson (1942), in response to Woolley's publication, explains that Woolley's line is "nothing other than Frisch's diagonal regression and is a statistical parameter, which has long appeared in literature. In terms of correlation surface it represents the major axis of the concentric ellipses of equal frequency". It is not difficult to see that Jones (1937), when describing the invariant regression introduced by Roos (1937) and Samuelson (1942), when noting the diagonal regression introduced by Frisch (1928), are both describing the same line. According to Jones (1937), the dimensions of the rectangle circumscribing the correlation ellipsoid are $2\sigma_x$ in the *x* direction and $2\sigma_y$ in *y* direction. One can now see that the slope of the diagonal of this rectangle is $\frac{\sigma_y}{\sigma_x}$. If we circumscribe rectangles around each of the concentric ellipses of equal frequency, it is apparent that they are all going to have the same diagonal as all the concentric ellipses (Figure 3).

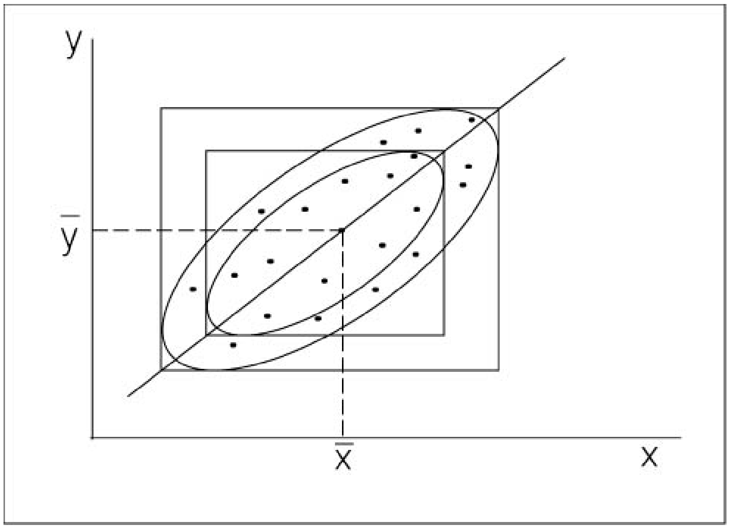

FIGURE 3

So we see that various people have proposed estimating procedures to build invariant lines and planes. Nevertheless, they all in general lead to the same estimators and coincide with the major axis of the concentric ellipses of equal frequency, as has been suggested by Samuelson (1942). A very early work in this direction is that of Ragnar Frisch (1928): *"Correlation and Scatter in statistical variables"*. However, for some reason, diagonal regression did not become popular, and since then has been rediscovered many times and in some cases extended by e.g. Teissier (1948), Barker et al. (1988), Draper and Yang (1997), Tofallis (2002).

Let us assume a sample of n observations and p variables, each observation is represented by a point X_i in p dimensional space. The sample is written as $(n \times p)$ data matrix X:

$$X = \{x_{ij}\} = \begin{array}{c} \textit{variables} \\ \begin{bmatrix} x_{11} \;\ldots\; x_{ip} \\ \vdots \quad x_{ij} \quad \vdots \\ x_{n1} \;\ldots\; x_{np} \end{bmatrix} \end{array} \textit{observations}$$

The rows of X standing for observations will be written $X_i = (x_{11}, x_{i2}, ..., x_{ip})$ and the columns standing for variables will be written $x_j = (x_{1j}, x_{2j}, ..., x_{nj})'$ we may write $X = (X_1, X_2, ..., X_n)' = (x_1, x_2, ..., x_p)$. For simplicity and without loss of generality we can assume that all variables are measured from their mean, thus $\mu_i = 0$. The product moments, taken about the means, are defined as: $M = X'X$. The moment matrix is:

$$M = \{m_{ij}\} = \begin{bmatrix} m_{11} \;\ldots\; m_{ip} \\ \vdots \quad m_{ij} \quad \vdots \\ m_{p1} \;\ldots\; m_{pp} \end{bmatrix}$$

The i^{th} diagonal element m_{ii} of M is the sum of squares of the variable x_i, $m_{ii} = \sum x_i x_i = x_i' x_i$ and $\sigma_i = +\sqrt{m_{ii}/n}$ is the standard deviation of x_i. Due to the symmetry property of the covariances, this is necessarily a symmetric matrix and positive definite (semi definite), self adjoint.

In p dimensions, the coefficients a_p of Frisch's invariant regression (in Frisch's notation) are given by solving the following eigen-system of equations, there is one equation for each value of i *, and i = 1...p*:

$$\sum_p (m_{ip} - \lambda_i m_{ii} e_{ip}) a_p = 0 \qquad (1),$$

where m_{ip} and m_{ii} are elements of the moment matrix, λ_i are characteristic roots and $e_{ip} = \begin{cases} 0\, when\ i \neq p \\ 1\, when\ i = p \end{cases}$. The coefficients a_k of the diagonal regression will correspond to the largest characteristic root λ_i satisfying the system of the equations.

Using the same notation, orthogonal regression in p dimensions will be:

$$\sum\nolimits_p (m_{ip} - \lambda_i e_{ip}) a_p = 0 \qquad (2),$$

Where m_{ip} is the element of the moment matrix, λ_i are characteristic roots and $e_{ip} = \begin{cases} 0\, when\ i \neq p \\ 1\, when\ i = p \end{cases}$, as in diagonal regression the coefficients of orthogonal regression will correspond to the largest characteristic root.

The term "diagonal" arises from the fact that the absolute values of the regression coefficients can also be determined by the square roots of the diagonal elements in the adjoint of the moment matrix. Frisch (1941) derives a general formula for the coefficients of diagonal regression (using Frisch's notation)

$$d_{ij} = \frac{\varepsilon_j}{\varepsilon_i} \sqrt{\frac{\hat{m}_{jj}}{\hat{m}_{ii}}} \qquad (3),$$

where d_{ij} are the diagonal regression coefficients, ε_j and ε_i are signs (1, −1 or 0), of $\hat{m}_{jj}$ and $\hat{m}_{ii}$ respectively, (these are the elements of the adjoint of the moment matrix). If we set $j = 1$ in (3) then the equation can be written as:

$$x_1 = \pm \frac{\sigma_1}{\sigma_2} x_2 \pm \frac{\sigma_1}{\sigma_3} x_3 \ldots \pm \frac{\sigma_1}{\sigma_p} x_p \qquad (4),$$

Cobb (1939) shows that an exceptional case occurs when the plane collapses to a line: in three dimensions it would take the form

$$\frac{x_1}{\sigma_1} = \frac{x_2}{\sigma_2} = \frac{x_3}{\sigma_3} = \frac{x_p}{\sigma_p}$$

An interesting feature of the diagonal regression line in two dimensions is that it is unique in being the only line-fitting technique that satisfies all of the following four properties:

1. For perfectly correlated variables the fitted line should reduce to the correct equation.
2. The fitted equation is invariant under an interchange of variables.

3. The regression is invariant under a simple dimensional or scale change in any of the variables.
4. The regression slope depends only upon the correlation coefficient and standard deviations.

These results were proved by another Nobel prize-winner, Paul Samuelson (1942). This set of properties motivates us to investigate the application of this method in multivariate statistics, particularly in Principal Component Analysis.

The problem of lack of invariance in PCA

To find the principal components one must solve the eigenvalue problem

$$(M - \lambda I)v = 0 \qquad (5),$$

Where *M* is the moment matrix, *I* is the identity matrix and λ is the eigenvalue. This problem is equivalent to finding numbers λ such that there is a nontrivial vector v with

$$Mv = \lambda I v \qquad (6),$$

The eigenvalues identify the size of the semi axes, and the eigenvectors give the directions of these axes (see Figure 4).

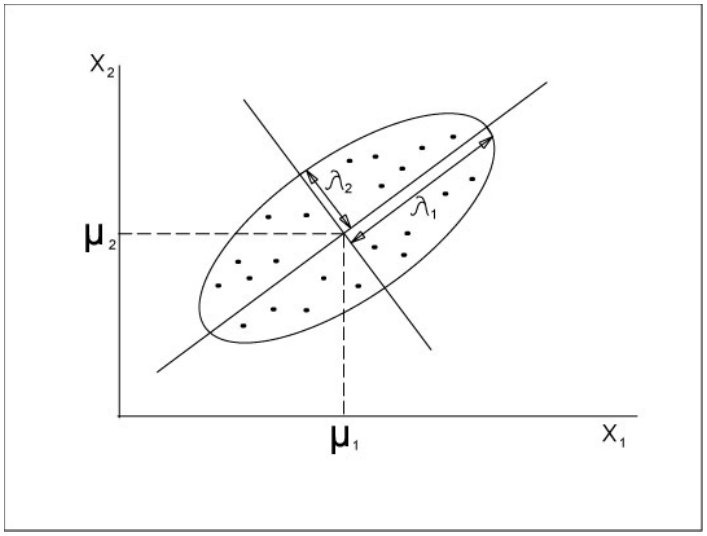

FIGURE 4

If equation (5) is to have a solution for *v* other than the zero vector then $(M - \lambda I)$ must be a non-singular matrix, thus it leads to the characteristic equation

$$\det(M - \lambda I) = 0 \qquad (7),$$

The determinant can be expanded to give a characteristic equation of n^{th} degree:

$$(-1)^n(\lambda^n + \alpha_{n-1}\lambda^{n-1} + \alpha_{n-2}\lambda^{n-2} + ... + \alpha_0) = 0 \qquad (8),$$

Equation (8) is called the characteristic equation of the matrix *M,* and the polynomial is called the characteristic polynomial. The roots of the characteristic equation are the eigenvalues. These *n* roots are non-negative since *M* is positive definite (semi definite). The sums of the squares of the original variables and of their principal components are the same.

$$trace(M) = \sum_{i=1}^{n} \lambda_i$$

Thus, we can say that each principal component accounts for a proportion

$$\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

of the overall variation in the original data.

Variables in the sample can be standardized by dividing by their standard deviation, in that case the moment matrix *M* becomes the correlation matrix *R.*

$$R = \{r_{ij}\} = \begin{bmatrix} r_{11} \; ... \; r_{1p} \\ \vdots \;\; r_{ij} \;\; \vdots \\ r_{p1} \; ... \; r_{pp} \end{bmatrix}$$

The diagonal elements r_{ii} of *R* are equal to *1* and r_{ij} is the correlation coefficient between x_i and x_j. If one uses the correlation matrix *R* instead of the moment matrix, the mathematical procedure for calculating the eigenvalues and eigenvectors is exactly the same. For the correlation matrix *R,* the diagonal elements are all unity. Hence, the sum of the variances of the standardized variables will equal *n,* which is the number of variables in the data set; so the proportion of variance acquired by the i^{th} principal component is simply λ_i / n. One can see from this that the eigenvalues and eigenvectors of the moment and correlation matrices are different and do not have a one-to-one relation.

Now let us assume that one of the variables has been multiplied by a scalar value *c.* To illustrate, let us take two centred variables y_1 and y_2 with moment matrix

$$K = \begin{bmatrix} y_1'y_1 & y_1'y_2 \\ y_2'y_1 & y_2'y_2 \end{bmatrix}$$

this leads to

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

and then the characteristic polynomial of the second degree is

$$\lambda^2 - \lambda(y_1'y_1 + y_2'y_2) + (y_1'y_1y_2'y_2 - y_1'y_2y_1'y_2) = 0$$

Let us change the scales of one of the variables, y_1 for instance, by multiplying by scalar *c*. The sum of squares of the y_1 will change from $y_1'y_1$ to $c^2y_1'y_1$. Therefore, the moment matrix on the left side of equation $Kv = \lambda Iv$ has changed, but the right hand side of the equation stays the same.

The new moment matrix K'' is

$$K'' = \begin{bmatrix} c^2y_1'y_1 & cy_1'y_2 \\ cy_2'y_1 & y_2'y_2 \end{bmatrix}$$

the characteristic polynomial is

$$\lambda^2 - \lambda c^2(y_1'y_1 + y_2'y_2) + c^2(y_1'y_1y_2'y_2 - y_1'y_2y_1'y_2) = 0$$

In both equations, the first member stays unchanged, but the second and the third members are different. In the case of more dimensions, it generalizes in an obvious way.

As result the roots of the two equations will not differ proportionally, thus we obtain different eigenvalues and eigenvectors that are not equivalent.

The literature offers two ways to solve this difficulty:

1. Use only variables measured in the same scales
2. Or, use the correlation matrix instead of the moment matrix.

PCA based on the correlation matrix will produce exactly the same eigenvalues and eigenvectors for both the scaled and unscaled data sets, however the solution will not reflect the variance structure of the variables and will stay the same as long as the correlation structure of the samples stays the same.

Invariant Principal Components

Frisch (1928) defines 'invariant regression' as being when the associated regression coefficient changes proportionally when one of the variables is rescaled. For instance, let us consider a regression equation for the relationship between price and quantity, where price is measured in pounds. For the particular quantity Q_1 units price equals P_1 pounds. Then, suppose the price axis is rescaled from pounds to pence. For the same quantity Q_1, the price will now be P_2 where $P_2 = 100P_1$ (see

Figure 5), the same price just measured in different units and therefore regression lines are invariant.

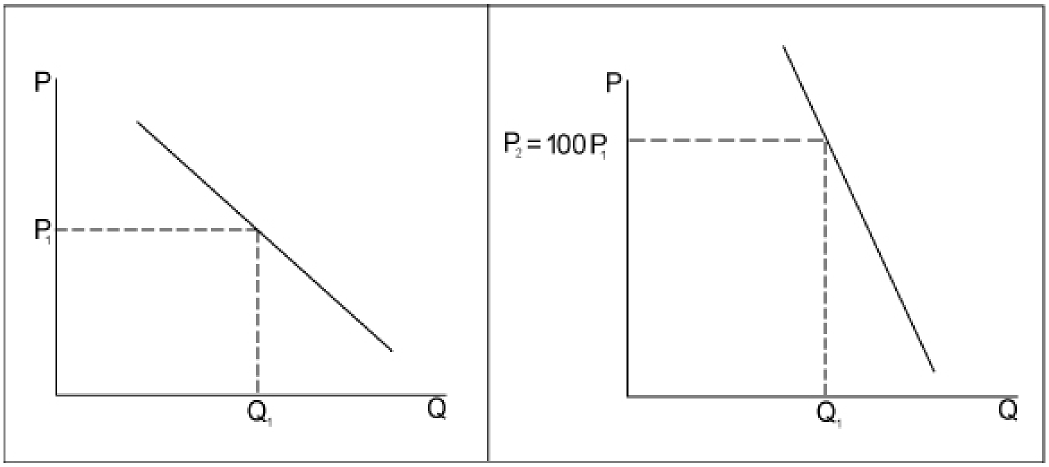

FIGURE 5

Frisch (1928) demonstrates that diagonal regression is invariant to change of scale of the original variables. Where diagonal regression is as represented by formulae (1), (3) and (4). One can compare formula (1) and formula (2) and see that the difference between them is the coefficient, (a sum of squares), attached to λ .

Now let us consider the situation where the identity matrix I in (5) is substituted by a diagonal matrix D, containing the products of the moments on the main diagonal. This matrix is defined as $d_{ii} = \sum x_i x_i = x_i' x_i$.

$$D = \{d_{ii}\} = \begin{bmatrix} d_{11} \; ... \; 0 \\ 0 \;\; ... \; d_{pp} \end{bmatrix}$$

Due to the properties of products of moments this matrix is positive definite and symmetric. The moment matrix M is as defined in the previous section.

Both matrices M and D are Hermitian, this is a consequence of them having only real entries and being symmetric $M^T = M$ and $D^T = D$. The moment matrix M is also positive definite (semi definite). Equation (1) leads to the generalized eigen-problem

$$(M - \lambda D)v = 0 \qquad (9),$$

Equations (1) and (9) are identical and are merely written using different notations. Hence, we introduce diagonal regression using a generalized eigen-problem approach, where:

$$M' = D^{-\frac{1}{2}} M D^{-\frac{1}{2}}$$

then the problem becomes:

$$M'v = \lambda v$$

As *D* is always positive definite M' is positive semi-definite. All eigenvalues of the definite pencil $\{M, D\}$ are real. This allows us to write them in sorted order $\lambda_1 \le \lambda_2 \le \lambda_3 \le ...\lambda_n$. If all $\lambda_i > 0$, then $M - \lambda D$ is called positive definite, and if all $\lambda_i \ge 0$, then $M - \lambda D$ is called positive semidefinite. Each eigenvector v_i is real, because *M* and *D* are real.

As in the previous section, let us consider the case of two variables, y_1 and y_2 with $\mu_1 = 0, \mu_2 = 0$. The moment matrix *L* is as defined as in the previous section and the diagonal matrix *D* is:

$$D = \begin{bmatrix} y_1'y_1 & 0 \\ 0 & y_2'y_2 \end{bmatrix}$$

The characteristic equation for this problem follows in the same way as equation (5), but instead of the identity matrix *I* we have matrix *D*.

$$\det(L - \lambda D) = 0$$

Expanding the determinant on the left hand side we have the following characteristic polynomial

$$y_1'y_1y_2'y_2(\lambda^2 - 2\lambda + 1) - y_1'y_2y_2'y_1 = 0 \qquad (10),$$

In equation (10) one can see that if one rescales one of the variables, then the whole equation changes proportionally. For example: change the scale of variable y_2 by multiplying it by scalar *c,* as we did in the previous section, equation (10) changes thus:

$$c^2(y_1'y_1y_2'y_2(\lambda^2 - 2\lambda + 1) - y_1'y_2y_2'y_1) = 0$$

Consequently we shall obtain the same roots (eigenvalues) and the eigenvectors will change proportionally. From the properties of principal components we know that “The sum of the squared correlations for each column equals the associated latent root, the amount of variance explained” Dunteman (1989). Hence the amount of variance explained by each component will not change either, but the eigenvector elements will change proportionally to reflect the changes in the variances of the original variables. One can see that the IPCA possesses the properties we require, namely: (i) the proportion of the overall variance explained by each component stays the same after the data set is rescaled, and (ii) the eigenvectors change proportionally according to the changes in the data set.

Numerical example

In this section we illustrate Invariant Principal Component Analysis using a “toy” example. The purpose is to demonstrate the properties of our analysis. We create a small data set containing three variables x_1, x_2, x_3 and calculate invariant components. Then we change the scales of the first two variables by multiplying through by ten and re-calculate the components. We denote the rescaled variables by y_1, y_2, y_3.

x_1	x_2	x_3	y_1	y_2	y_3
5.944	5.706	17.832	59.44	57.06	17.832
1.189	1.664	8.797	11.89	16.64	8.797
5.231	6.895	16.167	52.31	68.95	16.167
4.517	5.231	13.552	45.17	52.31	13.552
7.370	7.133	19.020	73.70	71.33	19.020
5.468	6.419	16.405	54.68	64.19	16.405
4.755	3.804	9.510	47.55	38.04	9.510
2.378	2.615	7.846	23.78	26.15	7.846
3.566	3.804	8.321	35.66	38.04	8.321
2.615	2.140	6.419	26.15	21.40	6.419

TABLE 1

Descriptive statistics for the original set are: $\mu_{x_1} = 4.30, \mu_{x_2} = 4.54, \mu_{x_3} = 12.39$, $\sigma_{x_1} = 1.87, \sigma_{x_2} = 2.01, \sigma_{x_3} = 4.71$, $\sigma_{x_1}^2 = 3.49, \sigma_{x_2}^2 = 4.05, \sigma_{x_3}^2 = 22.15$, and for rescaled set are: $\mu_{y_1} = 43.03, \mu_{y_2} = 45.41, \mu_{y_3} = 12.39$, $\sigma_{y_1} = 18.67, \sigma_{y_2} = 20.13, \sigma_{y_3} = 4.71$, $\sigma_{y_1}^2 = 348.52$, $\sigma_{y_2}^2 = 405.05$, $\sigma_{y_3}^2 = 22.15$.

Conventionally, one would standardize the data in both sets and perform PCA using correlation matrices. In that case of course we shall obtain the same results for both datasets: identical eigenvalues and eigenvectors. We shall take the data as displayed and only subtract the respective means from each variable. Table 2 shows the correlation between variables. Obviously the variables of the rescaled set have the same correlation structure.

	x_1	x_2	x_3
x_1	1.00		
x_2	0.93	1.00	
x_3	0.88	0.93	1.00

TABLE 2

The methodology of PCA suggests that only if variables in the data set are correlated is there any point in proceeding with such an analysis. Using our proposed method, solving (9) we obtain the following results for the original dataset. (Computations were carried out using Matlab's built-in *eig* function.)

Eigenvalues are: $\lambda_1 = 2.8228,\ \lambda_2 = 0.1224,\ \lambda_3 = 0.0549$ and eigenvectors are given in Table 3:

	PC1	PC2	PC3
x_1	0.10237	-0.12893	-0.069096
x_2	0.096787	0.0048592	0.13433
x_3	0.040673	0.048947	-0.031077

TABLE 3 Principal Components calculated using the diagonal regression approach

For the rescaled data set the eigenvalues are;

$\lambda_1 = 2.8228,\ \lambda_2 = 0.1224,\ \lambda_3 = 0.0549$ (same as above), and eigenvectors are:

	PC1	PC2	PC3
y_1	0.010238	-0.012892	-0.0069131
y_2	0.009678	0.00048259	0.013432
y_3	0.040673	0.048955	-0.031064

TABLE 4

As expected we obtain the same eigenvalues, for both datasets; this can be explained by the fact that the correlation structure has not changed with rescaling of the variables. The variance structure has changed however, and we obtain proportionally adjusted eigenvector elements identifying the new directions, in accordance with the change in scales. This can be seen from the ratios between the eigenvectors. One can divide components associated with, for instance, the third variable in each data set by components associated with the first and the second variables and see that the ratios have changed proportionally (results are given in Table 5).

	PC1	PC2	PC3
x_3 / x_1	0.39731367	-0.37964011	0.449766
y_3 / y_1	3.97274858	-3.79731617	4.493498
x_3 / x_2	0.42023206	10.0730573	-0.23135
y_3 / y_2	4.20262451	101.442218	-2.31269

TABLE 5 Note that the first two rows differ by a factor of 10, as do the last two rows.

Table 6 shows the squared correlations (these are invariant to change of scale, and so are the same for both scaled and unscaled data sets); note how high these are for the first component using IPCA.

	PC1	PC2	PC3
x_1	0.9279	0.0637	0.00821
x_2	0.9638	0.0001	0.0360
x_3	0.9309	0.0584	0.0105
Sum	2.8226	0.1222	0.05471

TABLE 6 Squares of correlations between principal components and variables.

Summing each column in Table 6, we see that sum of squared correlations equals the corresponding eigenvalue for the principal component; and the sum of the eigenvalues equals the number of the variables. Hence, the proportion of the overall variance that each principal component explains can be calculated according to the same formula as used in conventional PCA based on the correlation matrix:

$$\frac{\lambda_i}{n}$$

The first component explains 95% of overall variance, the second 4% and the third explains only 1%.

Table 7 shows the principal components obtained using traditional PCA based on the correlation matrix

PC1	PC2	PC3
-0.573364	0.722079	0.387112
-0.584331	-0.029208	-0.81099
-0.574292	-0.691194	0.438679

TABLE 7

These components are the same for both the rescaled and unscaled datasets, from which fact it follows that they do not reflect the variance structure of the data. Calculated eigenvalues are as follows: $\lambda_1 = 2.8228,\ \lambda_2 = 0.1224,\ \lambda_3 = 0.0549$. One notices that traditional PCA based on the correlation matrix has the same eigenvalues as our Invariant PCA. Thus, one can see that IPCA reflects the correlation structure of the data in the same way, but in addition, the eigenvectors describe variance structure of the datasets unlike PCA based on correlation matrix.

Note that the calculated eigenvectors are not normalized, but as Cadima. J. and Jolliffe (1997) point out: "The different scalings change the size of the vector but not its direction. Relative values of loadings in the vector are unchanged". By definition, elements of the first eigenvector are coefficients of the best fitting plane. Hence they have to be the same as the coefficients from Frisch's diagonal regression (4).

Setting the first principal component (PC1) to zero:

$$0.10237x_1 + 0.096787x_2 + 0.040673x_3 = 0 \qquad (11),$$

and re-arranging:

$$x_1 = -(0.93x_2 + 0.397x_3) \qquad (12),$$

One can substitute values in the formula (4) and see that they are identical to the coefficients in (12). Likewise, we get identical results for the rescaled dataset.

Conclusion

We have presented the application of diagonal regression to Principal Component Analysis. The problem has been introduced using generalized eigen-analysis. The use of diagonal regression allows us to build scale-independent (i.e. unit-invariant) models which reflect not only the correlation structure of the data set, but the variance structure as well. This combination of properties is not shared by traditional PCA based on the correlation matrix. The invariant results of PCA based on the correlation matrix reflect the correlation structure but not the variance structure of the data, as standardized variables all have variance equal to unity, i.e. all the variables are assumed equally important.

A numerical example was employed to illustrate some properties of Invariant Principal Component Analysis. The correlation between the components and the variables of the rescaled and original datasets were shown to be the same and eigenvectors differ accordingly. This property illustrates that the results arising from Invariant PCA are scale-independent with coefficients which reflect the variance structure of the data.

Reference

Adcock, R. J., 1877. "A Problem in Least Squares". *Analyst,* **4,** 183.

Ahamad, B., 1967. "An analysis of crimes by the method of principal components". *Applied statistics,* **16,** 17-35.

Barker, F., Soh, Y. C. and Evans R.J. 1988. "Properties of the geometric mean functional relationship". *Biometrics,* **44,** 279-281.

Cadima. J. and Jolliffe, I. T., 1997. "Some comments on ten Berge, J.M.F. & Kiers, H.A.L. (1996). Optimality criteria for principal component analysis and generalizations". *British Journal of Mathematical and Statistical Psychology,* **50,** 365-366.

Campbell, N. A., 1980. "Robust procedures in Multivariate analysis: Robust Covariance Estimation". *Applied statistics,* **29,** 231-237.

Chatfield, C. and Collins, A., J., 1986. *Introduction to Multivariate statistics.* Cambridge:The University Press

Cobb, C. W., 1939. "Note on Frisch's Diagonal Regression". *Econometrica,* **7,** 77-80.

Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R., 1981. "Robust estimation of dispertion matrices and Principal Components". *Journal of the American Statistical Assosiation,* **76,** 354-362.

Draper, N. R. and Yang, Y. F., 1997. "Generalization of the geometric mean functional relationship". *Computational statistics & data Analysis,* **23,** 355-372.

Dunteman, G. H., 1989. *Principal Components Analysis.* Sara Miller McCune, Sage Publications, Inc

Frisch, R., 1928 Correlation and scatter in statistical variables In *Foundations of modern econometrics: selected essays of Ragnar Frisch (1995)*, Vol. 1 (Ed, Bjerkholt, O.) Edward Elgar, London.

Frisch, R., 1934. *Statistical Confluence Analysis by Means of Complete Regression Systems.* Oslo:

Frisch, R., 1941. "Editorial Notes". *Econometrica,* **9,** 94-95.

Gnanadesikan, R., 1977. *Methods for statistical data analysis of multivariate observations.* John Willey & Sons

Healy, M. J. R., 2000. *Matrices for statistics,* 2nd edition. Oxford:Clarendon Press

Hotelling, H., 1933. "Analysis of a Complex Statistical Variables into Principal Components". *Journal of Educational Psychology***,** 417-441, 498-520.

Jackson, J. E., 2003. *A user's guide to principal components.* A John Wiley & Sons

Jolliffe, I. T., 2002. *Principal Component Analysis (Springer series in statistics).* Second edition. Springer-Verlag

Jones, H. E., 1937. "Some Geometrical Considerations in the General Theory of Fitting Lines and Planes". *Metron,* **13,** 21-30.

Kaiser, H. F., 1967. "Uncorrelated linear composites maximally related to a complex of correlated observations". *Educational and Psychological Measurement***,** 3-6.

Kendall, M. G., 1965. *A Course in Multivariate Analysis.* Third Impression. London: Charles Griffin & Company Limited

Korhonen, P., J., 1984. "Subjective Principal Component Analysis". *Computational Statistics & Data Analysis,* **2,** 243-255.

Lardic, S., Priaulet, P. and Priaulet, S., 2002. "PCA of yield curve dynamics: Questions of methodologies". *Journal of bond trading and management,* **1,** 327-349.

Loretan, M., 1997. "Generating market risk scenarios using principal component analysis: methodological and practical considerations". *Federal reserve board.*

Mathews, J. N. S., 1984. "Robust methods in the Assesment of Multivariate normality". *Applied statistics,* **33,** 272-277.

Meredith, W. and Millsap, E. R., 1985. "On Component Analysis". *Psychometrika,* **50,** 495-507.

Pearson, K., 1901. "On lines and planes of the closest fit". *The philosophical magazine,* **2,** 559-572.

Rao, C. R., 1952. *Advanced statistical methods in biometric research.* Wiley

Reed, L. J., 1921. "Fitting Straight Lines". *Metron,* **1,** 54-61.

Roos, C. F., 1937. "A general invariant criterion of fit for lines and planes where all variates are subject to error". *Metron,* **13,** 3-20.

Samuelson, P., 1942. "A Note on Alternative Regression". *Econometrica,* **10,** 80-83.

Sokal, R. R. and Rohlf, F. J., 1981. *Biometry: the principles and practice of statistics in biological research.* Third Edition. W.H. Freeman and Company

Teissier, G., 1948. "La relation d'allometrie: Sa signification statistique et biologique". *Biometrics,* **4,** 14-48.

ten Berge, J. M. F. and Kiers, H. A. L., 1996. "Optimality criteria for principal components analysis and generalizations". *British Journal of Mathematical and Statistical Psychology,* **49,** 335-345.

Tofallis, C., 2002 Model Fitting for Multiple Variables by Minimizing the Geometric Mean Deviation. In *Total Least Squares and Errors - in- Variables Modeling: Algorithms, Analysis and Application* (Eds, Van Huffel, S. and Lemmerling, P.) Kluwer Academic, Dordrecht.

Wald, A., 1940. "The Fitting of Straight Lines if Both Variables Subject to Error". *The Annals of Mathematical Statistics,* **11,** 284-300.

Walker, A. M., 1967. "Some critical comments on "An analysis of crimes by the method of Principal Components" by B. Ahamad". *Applied statistics,* **16,** 36-39.

Woolley, E., 1941. "The Method of Minimized Areas as a Basis for Correlation Analysis". *Econometrica,* **9,** 38-62.