Some Thoughts on Understanding Correlation Matrices

Maryam Hadavand-Siri and Clayton V.Deutsch

The correlation matrix is a positive semi definite matrix that describes the dependency between different data sets. In case of multi variables mode or dealing with many secondary variables which is hard to predict spatial distribution and dependency between variables, correlation matrix is a key element to describe this dependency. The principal directions of data set variance are defined by principal components. Principal Component Analysis (PCA) is a statistical procedure to calculate eigenvalues and eigenvectors of correlation matrix which are principal component of data set, by dimension reduction.

Introduction

Dependency refers to any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationship involving dependence. Correlation between two set of data set X, Y defined as:

$$\rho_{XY} = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y]}{\sigma_X \sigma_Y} \quad (1)$$

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y) \quad (2)$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu_X)}{n}} \qquad \sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\mu_Y)}{n}} \quad (3)$$

Where $\rho, E\ and\ Cov$, are correlation, expected value and covariance operator respectively, μ is the mean, σ is standard deviation and n is number of variables. The correlation is +1 in the case of a perfect positive linear relationship, −1 in the case of a perfect negative linear relationship, and some value between −1 and 1 in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship (closer to uncorrelated). The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables. The correlation cannot exceed 1 in absolute value. When correlations among several variables are computed, they are typically summarized in the form of a correlation matrix. Correlation matrices are built to describe the dependency between different data sets and are symmetric as $Corr(X,Y) = Corr(Y,X)$. Correlation matrices must be positive semi definite. It means for all Non-zero column vector Z, $Z^T \rho Z > 0$ (Z^T is the transpose of Z and ρ is correlation matrix).

The subject of multivariate analysis deals with the statistical analysis of the data collected on more than one variable. These variables may be correlated with each other, and their statistical dependence is often taken into account when analyzing such data. Correlation matrix is a key element to explain and apply this dependency in multi variable mode. In reservoir estimation the primary well data, which is expensive to obtain by drilling is predicted using the easy and cheap obtaining secondary seismic data. Correlation matrix can be useful for spatial prediction and dimension reduction when we are dealing with many secondary variables (Kumar, A. and Deutsch, C.V., 2009, CCG annual report).

Eigenvalues and eigenvectors:

Principal component of a data set are found by calculating the eigenvalues and eigenvectors of the data covariance matrix. In fact, eigenvalues are the variance of principal components. Suppose that A is a square matrix of size n, $X \neq 0$ is a vector in C^n, and λ is a scalar in C. Then X is an eigenvector of A with eigenvalue λ, if $AX = \lambda X$. The eigenvalues of a matrix *A* are precisely the solutions λ to the characteristic equation (I is identity matrix).

$$\det(A - \lambda \mathrm{I}) = 0$$

$$\det(A-\lambda \mathrm{I}) = \det\left(\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & a_{nn} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right) \quad (4)$$

$$= det\begin{bmatrix} a_{11}-\lambda & 0 & \cdots & 0 \\ 0 & a_{22}-\lambda & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & a_{nn}-\lambda \end{bmatrix}$$
$$= (a_{11}-\lambda)(a_{22}-\lambda)\ldots(a_{nn}-\lambda) = 0$$

The solutions to this equation are the eigenvalues $\lambda_i = a_{ii}$ (i=1,2,…n). When we deal with large size of matrices, get the eigenvalues and eigenvectors from characteristic equation is not an easy job. There are two classes of numerical methods to calculate eigenvalues and eigenvectors (Panhuis, P.H.M., 2005): (1) Partial methods - computation of extrema eigenvalues such as power method, and (2) Global methods - approximation of whole spectrum such as: Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), and Factorization. This report provides an introduction to global methods and specifically focuses on PCA method. MDS and factorization will be covered in future works.

Principal Component Analysis (PCA):

Principal Component Analysis, or simply PCA, is a statistical procedure concerned with elucidating the covariance structure of a set of variables. In particular it allows us to identify the principal directions in which the data varies. Principal Component Analysis (PCA) is a method of identifying the pattern of data set by a much smaller number of "new" variables, named as principal components (Gillies,D.,,2005). For example, in figure 1, suppose that the small circles represent a two variable data set which we have measured in the X-Y coordinate system.

Red Circles: $x_1, x_2, \ldots x_n$ $Mean=\mu_x$

Blue Circle $y_1, y_2, \ldots, y_n$ $Mean=\mu_y$

The principal direction in which the data varies is shown by the U axis and the second most important direction is the V axis orthogonal to it. If we place the [U,V] axis system at the mean of the data(μ_x, μ_y) it gives us a compact representation. If we transform each [X,Y] coordinate into its corresponding [U,V] value, the data is de-correlated, meaning that the co-variance between the U and V variables is zero. For a given set of data, principal component analysis finds the axis system defined by the principal directions of variance (ie the [U,V] axis system in figure 1) . The directions U and V are called the principal components.

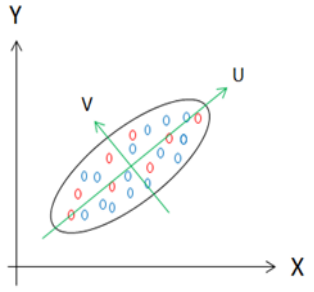

Figure 1

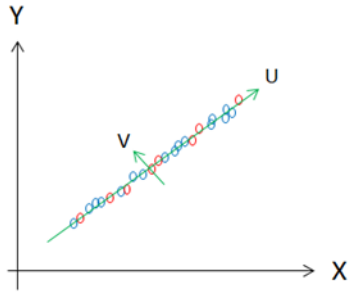

Figure 2

If the variation in a data set is caused by some natural property, or is caused by random experimental error, then we may expect it to be normally distributed. In this case we show the nominal extent of the normal distribution by a hyper-ellipse (the two dimensional ellipse inFigure1). The hyper ellipse encloses data points that are thought of as belonging to a class. It is drawn at a distance beyond which the probability of a point belonging to the class is low, and can be thought of as a class boundary.

If the variation in the data is caused by some other relationship, then PCA gives us a way of reducing the dimensionality of a data set. Consider two variables that are nearly related linearly as shown in figure2. As in figure1 the principal direction in which the data varies is shown by the U axis and the secondary direction by the V axis. However in this case all the V coordinates are very close to zero. We may assume, for example, that they are only non-zero because of experimental noise or measurement

error. Thus in the U-V axis system we can represent the data set by one variable U and discard V. Thus we have reduced the dimensionality of the problem by 1.

In computational terms the principal components are found by calculating the eigenvectors and eigenvalues of the data covariance matrix. This process is equivalent to finding the axis system in which the co-variance matrix is diagonal. The eigenvector with the largest eigenvalue is the direction of greatest variation, the one with the second largest eigenvalue is the (orthogonal) direction with the next highest variation and so on. In the other word, eigenvalues are the variance of principal components. The first eigenvalue is the variance of the first principal component; the second eigenvalue is the variance of the second principal component and so on (Gillies,D., 2005).

The eigenvalues of A, $n \times n$ matrix, are defined as the roots of:

$$\det(A - \lambda) = |A - \lambda \mathrm{I}| = 0 \quad (5)$$

Let λ be an eigenvalue of A. Then there exists a vector x such that:

$$Ax = \lambda x \quad (6)$$

The vector x is called an eigenvector of A, associated with the eigenvalue λ. Notice that there is no unique solution for x in the above equation. It is a direction vector only and can be scaled to any magnitude. To find a numerical solution for x we need to set one of its elements to an arbitrary value, say 1, which gives us a set of simultaneous equations to solve for the other elements. If there is no solution we repeat the process with another element. Ordinarily we normalize the final values so that x has length one, that is $x.x^T = 1$.

Suppose we have a 3×3 matrix A with eigenvectors x_1, x_2, x_3 and eigenvalues $\lambda_1, \lambda_2, \lambda_3$ so:

$$Ax_1 = \lambda_1 x_1 \qquad Ax_2 = \lambda_2 x_2 \qquad Ax_3 = \lambda_3 x_3 \quad (7)$$

Putting the eigenvectors as the columns of a matrix gives:

$$A[x_1 \quad x_2 \quad x_3] = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} [x_1 \quad x_2 \quad x_3] \quad (8)$$

Writing:

$$\varphi = [x_1 \quad x_2 \quad x_3] \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \quad (9)$$

Gives us the matrix equation:

$$A\varphi = \Lambda\varphi \quad (10)$$

We normalized the eigenvectors to unit magnitude, and they are orthogonal, so:

$$\varphi\varphi^T = \varphi^T\varphi = I \quad (5)$$

This means that:

$$\varphi^T A \varphi = \Lambda \quad (6)$$

And:

$$A = \varphi\Lambda\varphi^T \quad (7)$$

Now let us consider how this applies to the covariance matrix in the PCA process. Let Σ be a $n \times n$ covariance matrix. There is an orthogonal $n \times n$ matrix φ whose columns are eigenvectors of Σ and a diagonal matrix Λ whose diagonal elements are the eigenvalues of Σ, such that:

$$\varphi\Sigma\varphi^T = \Lambda \quad (8)$$

We can look on the matrix of eigenvectors φ as a linear transformation which, in the example of figure1 transforms data points in the [X, Y] axis system into the [U,V] axis system. In the general case the linear transformation given by φ transforms the data points into a data set where the variables are uncorrelated. The correlation matrix of the data in the new coordinate system is Λ which has zeros in all the off diagonal elements.

Principal component analysis is appropriate when you have obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. Some limitations of PCA (Izenman, A.J.,2008): (1) The directions with largest variance are assumed to be of most interest. (2) We only consider orthogonal transformations (rotations) of the original variables. (Kernel PCA is an extension of PCA that allows non-linear mappings). (3) PCA is based only on the mean vector and the covariance matrix of the data. Some distributions (e.g. multivariate normal) are completely characterized by this, but

others are not. (4) Dimension reduction can only be achieved if the original variables were correlated. If the original variables were uncorrelated, PCA does nothing, except for ordering them according to their variance. (5) PCA is not scale invariant.

Implementation

In this study, a program written in fortran90 code used to calculate correlation matrix, eigenvalues and eigenvectors for a data set with six variables. This code reads data set in ASCII format, deletes null values (use full valued subset to delete -999) and it calculates Mean, Standard deviation, Covariance and Correlation matrix for data set. Then calculate eigenvalues and eigenvectors for the correlation matrix. Standard GSLIB convention (corrmat_plot) is used to plot correlation matrix.

Calculated correlation matrix for six different variables displayed in Figure 3. This is an appropriate opportunity to review just how a correlation matrix is interpreted. The rows and columns of Figure 3 correspond to the six variables included in the analysis: Row 1 (and column 1) represents variable 1, row 2 (and column 2) represents variable 2, and so forth. When a given row and column intersect, you will find the correlation between the two corresponding variables. For example, where the row for variable 2 intersects with the column for variable 1, you find a correlation of 0.14; this means that the correlation between variables 1 and 2 is 0.14.

Based on Figure 3, variables 2, 5 and 6 show relatively strong positive correlation with one another $(\rho_{25=}0.76\,, \rho_{26=}0.66, \rho_{56=}0.56)$.

Variable 4 shows relatively strong negative correlation with variable 2, 5 and 6 $(\rho_{42=} - 0.73\,, \rho_{45=} - 0.59, \rho_{46=} - 0.46)$.

Variable 3 also shows negative correlation with variable 2, 5 and 6 $(\rho_{32=} - 0.53\,, \rho_{35=} - 0.38,\ \rho_{36=} - 0.49)$.

However, variable 1 has no correlation with the other variables. When the correlation between two variables is less than 0.2 ($|\rho| < 0.2$) we assume they are uncorrelated (Babak, O. and Deutsch, C.V.,2008). Let's reorder variables based on their correlations to each other and have a new correlation matrix for reordered variables. Variable 2, 5 and 6 which have strong positive correlation to each other place in first, second and third orders respectively. Variable 4 and 3 which have negative correlation with variables 2, 5 and 6 place in fourth and fifth order respectively. Variable 1 has no correlation with the other variables places at the last order.

1	Ni(V1)	Fe(V2)
2	Fe(V2)	Co(V5)
3	SiO2(V3)	Al2O3(V6)
4	MgO(V4)	MgO(V4)
5	Co(V5)	SiO2(V3)
6	Al2O3(V6)	Ni(V1)

Table1: Reordering variables based on their correlation

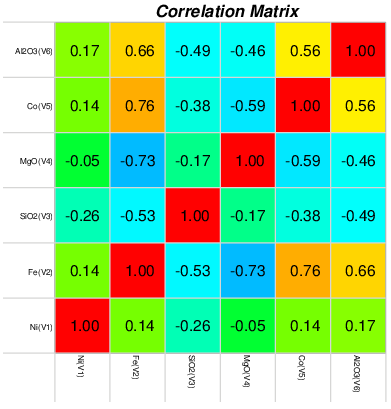

Figure 3: Correlation matrix for six different variables

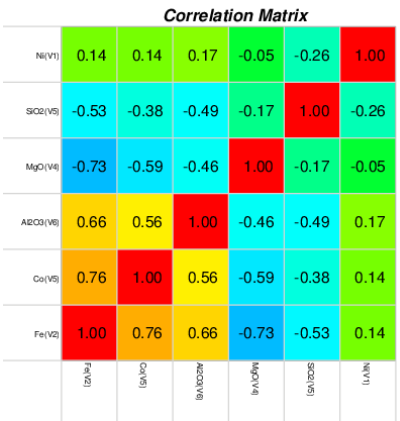

Figure 4: Correlation matrix based on reordered variables

The new correlation matrix (Figure4) shows that the six variables seem to hang together in three distinct groups. First group, variable2,5 and 6 which they have strong positive correlation to each other. Second group, variable3 and 4 which they have negative correlations with group1. The last group is variable1 has no correlation with the rest of variables (Figure 5). This is the redundancy of six variables to three. In essence, this is what accomplished by correlation matrix. In multivariate analysis mode or when dealing with many secondary variables, correlation matrix allows you to reduce a set of observed variables into a smaller set of artificial variables which called dimension reduction.

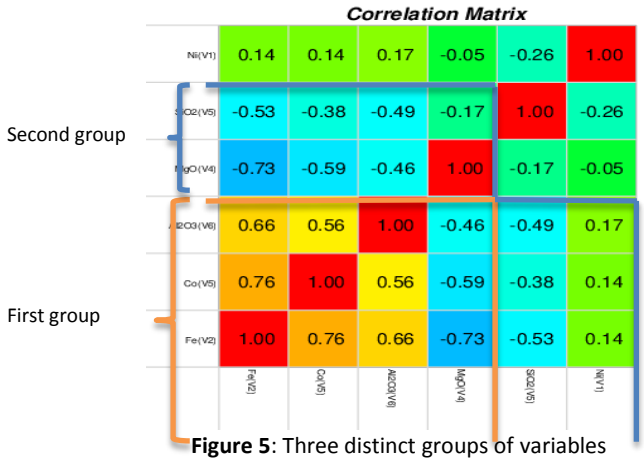

Figure 5: Three distinct groups of variables

Eigenvalues and eigenvectors are calculated for this correlation matrix and displayed on Table2. Then they are reordered based on their magnitudes as the first eigenvector with largest eigenvalue is the direction of greatest principal component and so on (Table 3).

	Eigenvalues					
	0.009	3.1376	0.8583	1.2835	0.2816	0.43
	Eigenvector					
Ni(V1)	-0.0553	0.1469	0.8668	0.4732	0.0095	0.0025
Fe(V2)	-0.6259	0.5325	-0.083	-0.0748	-0.5228	-0.1971
SiO2(V3)	-0.4957	-0.3081	0.3746	-0.6546	0.2518	0.1645
MgO(V4)	-0.5912	-0.3942	-0.2838	0.5675	0.2962	-0.0691
Co(V5)	0.0025	0.4836	-0.0306	-0.1053	0.7021	-0.511
Al2O3(V6)	-0.0994	0.4589	-0.1411	0.0944	0.2872	0.8174

Table 2: Eigenvalues and eigenvectors for correlation matrix

	Re-ordered eigenvalues					
	3.1376	1.2835	0.8583	0.43	0.2816	0.009
	Cumulative eigenvalues					
	0.52293	0.73685	0.8799	0.95157	0.9985	1
	Eigenvectors					
Ni(V1)	0.1469	0.4732	0.8668	0.0025	0.0095	-0.0553
Fe(V2)	0.5325	-0.0748	-0.083	-0.1971	-0.5228	-0.6259
SiO2(V3)	-0.3081	-0.6546	0.3746	0.1645	0.2518	-0.4957
MgO(V4)	-0.3942	0.5675	-0.2838	-0.0691	0.2962	-0.5912
Co(V5)	0.4836	-0.1053	-0.0306	-0.511	0.7021	0.0025
Al2O3(V6)	0.4589	0.0944	-0.1411	0.8174	0.2872	-0.0994

Table 3: Reordered eigenvalues and eigenvectors for correlation matrix

Cumulative eigenvalue curve c(x) is defined to be:

$$C(x) = \frac{\sum_i^x \lambda_i}{\sum_{i=1}^N \lambda_i} \times 100 \qquad (15)$$

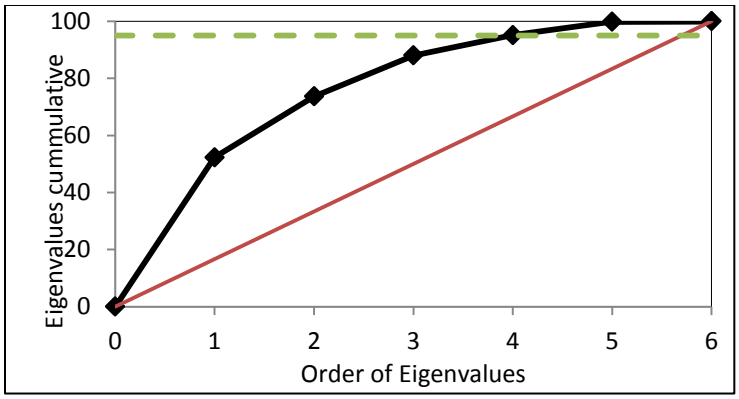

Figure 6: Cumulative eigenvalue VS Order of eigenvalues

The interpretation of this curve, Figure 6, is that the value C(x) represents the amount of information maintained in the input vectors if we project them onto the subspace spanned by the top x eigenvectors. A feature transformation that, for instance, retains 50 percent of the original information (variance) of the input data can be obtained by first eigenvalue(C(x) =50), and more than 95 percent of the original information can be obtained by first four eigenvalues (C(x) =95). If the data are independent then cumulative curve, C(X), follows the red line. Consider the following classification for the eigenvectors:

$$if \begin{cases} 0.2 \le \rho \le 1 & positive\ correlation \rightarrow Green \\ |\rho| < 0.2 & no\ correlation \rightarrow Gray \\ -1 \le \rho \le -0.2 & negative\ correlation \rightarrow Orange \end{cases}$$

Then, it is easier to visualize and interpret eigenvectors based on variables correlation (Table 4). First eigenvectors (λ=3.1376) shows strong positive correlation for variable 2, 5 and 6, negative correlation for variable 3 and 4 and de-correlation for variable 2 as expected.

	Re-ordered eigenvalues					
	3.1376	1.2835	0.8583	0.43	0.2816	0.009
	Cumulative eigenvalues/Number of variables					
	0.52293	0.73685	0.8799	0.95157	0.9985	1
	Eigenvectors					
Fe(V2)	0.5325	-0.0748	-0.083	-0.1971	-0.5228	-0.6259
Co(V5)	0.4836	-0.1053	-0.0306	-0.511	0.7021	0.0025
Al2O3(V6)	0.4589	0.0944	-0.1411	0.8174	0.2872	-0.0994
MgO(V4)	-0.3942	0.5675	-0.2838	-0.0691	0.2962	-0.5912
SiO2(V3)	-0.3081	-0.6546	0.3746	0.1645	0.2518	-0.4957
Ni(V1)	0.1469	0.4732	0.8668	0.0025	0.0095	-0.0553

Table 4: Reordering eigenvectors based on variable correlation (Table1)

Summary and Future work

The correlation matrix summarizes correlation among several variables can describe statistical dependency between them. When dealing with large size of matrices, it is difficult to calculate eigenvalues and eigenvectors from characteristic equation. There are several numerical methods to calculate eigenvalues and eigenvectors for large size of matrices. Principal component analysis (PCA) calculates eigenvalues and eigenvectors by dimension reduction. In future work, other numerical methods such as Multi-Dimensional scaling and factorization will be described.

References

Babak, O., and Deutsch, C.,V., 2008, CCG annual report: Testing for the Multivariate Gaussian Distribution of Spatially Correlated Data.

Gillies, D., 2005, Lectures and course materials: Principal Component Analysis.

Izenman, A.J., 2008, Modern Multivariate Statistical Techniques, pp 597-606

Kumar, A., and Deutsch,C.V., 2009, CCG annual report :Optimal Correlation of Indefinite Correlation Matrices.

Panhuis, P.H.M.W., 2005, Iterative Techniques For Solving Eigenvalue problems.

Wickelmaier, F.,2003, An Introduction to MDS.