Geostatistical Analysis by Vector Space Methods

By

J.B. Olaleye (DR.) & J.O. Sangodina
Dept. of Geoinformatics and Surveying
Faculty of Engineering
University of Lagos
Akoka, Yaba.

Abstract
A large proportion of the operations in geoinformation processing has to do with decision making. Such decision making exercise is often based on the results of some statistical calculations involving stochastic variables. The modern geomatics engineer or GIS consultant must therefore be properly equipped with adequate statistical knowledge and processing skill to be able to summarize experimental data and draw useful conclusions..

However, because of the usually crowded training programmes of the geoinformation professional, not enough time is available for geostatistical studies. And due to the often long and tedious processes involved in the statistical calculations, he is not able to develop the right skill in this very vital area of his professional practice. Fortunately, the geomatics engineer is often well equipped with sufficient mathematical preparation especially in aspects of linear algebra and vector space methods.

The use of vector space methods in statistical analysis makes it easier to understand the concepts and applications of the basic methods of data summarization and information extraction. The major advantage being that of economy of the process, giving the user the freedom of mind to think of other more advanced uses of statistical concepts in geoinformation processing.

This paper presents the basic concepts of vector space methods and highlights their applications in the process of extracting or summarization of information from collected data. Computational examples show that indeed, statistical analysis procedure is not and should not be as tedious as it is often presented in the literature.

1.0 Introduction

A large proportion of the operations in geoinformation processing has to do with decision making. Information is a vital ingredient for sound decision. As a matter of fact, the importance of accurate, sufficient and timely information in any decision-making process cannot be over emphasized. The lack of information can only lead to wrong decisions whose effect may be disastrous. Information is an item of data, which is useful for a particular purpose. In other words, information is not just data, but data that is useful for some applications. In most cases, information is extracted from sampled data through some statistical calculations. Thus, most of the activities of the geomatics engineer involves the designing, collection, and analysis of experimental data in order to establish, confirm or disprove some hypothetical statements about invents and processes happening within his environment.

In practice, the educational programmes for the training of the professional geomatics engineer are often so crowded that there is hardly any opportunity for extra emphasis on any particular course. In most cases, statistics is taught as part of the year one mathematical course and not followed-up with a more practical course later in the programme. This invariably makes the student to feel that skills in statistical perception are not essential.

The use of vector space methods in statistical analysis is rather new. An excursion into statistical literature reveals that authors generally prefer the long hand approach of data summarization (Green and Carol, 1976). Although, results are obtained using this method, the tediousness of the algorithms poses enough problems to discourage a casual user. The introduction of vector space concepts into statistical calculations serves to make the formulations compact, intuitive and appealing to all users. The immediate effect of this is that the professional engineer is better able to understand the concepts and applications of the basic methods of data summarization and information extraction. The long-term effect is that of stimulating interest in statistical experimentation which enhances and broadens the horizon of professional practice.

In the rest of this paper, the basic concepts of linear vector space methods are presented. Specifically, the rudiments of vectors and matrices, which are directly useful in statistical analysis, are discussed in some detail. A sample geospatial problem is then posed and attempts are made to go through the various solution steps as a way of showing the beauty of the vector space approach.

2.0 The Hilbert Vector Space Concepts

A vector space is a linear space in which entities exist which may be represented by a set of numbers referred to a coordinate system. The concept of a vector space has emerged as a special case of the abstract set theory through the definition of a set whose elements are composed of a collection of real numbers referred to a Cartesian system of coordinates. There are many spaces whose membership criteria and the kinds of operations allowed on the are members may be defined. Of particular interest in this paper is the so called Hilbert space which is defined as a non-empty set in which all linear operations such as addition and scalar multiplication on its elements are permitted. Furthermore, the idea of scalar product (inner product) of any two members of the space is allowed. This means that the inner product of two elements of the space results in a real (scalar) number (Olaleye, 1992).

In general, vector space concepts are usually given practical meanings through the definitions and operations on entities called vectors and matrices. These are explained briefly in what follows.

2.1 Vector

A vector a of order nx1 is an ordered set of n real numbers (scalars), which may be written as

$$a = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_n \end{bmatrix}$$

Some special forms of vectors which are useful for statistical calculations are ***Null, Unit, Sign, and Zero-One Vectors.*** These are briefly summarised here.

If all components of a vector are zero, we shall call this *null* or zero vector, denoted as 0. This should not be confused with the scalar 0. If all components of a vector are 1, this type of vector is called a unit vector, denoted as 1. If the components consists of either 1's or -1's (with at least one of each type present), this is called a *sign* vector. If the components consist of either 1's or 0's (with at least one of each type present), this will be called a zero-one vector. Examples:

	Column vectors	Row vectors
Null vectors	$\begin{bmatrix}0\\0\\0\end{bmatrix}$	(0,0,0,0)
Unit vectors	$\begin{bmatrix}1\\1\end{bmatrix}$	(1,1,1)
Sign vector	$\begin{bmatrix}-1\\1\\-1\end{bmatrix}$	(1, -1)
Zero-one vectors	$\begin{bmatrix}0\\1\\1\end{bmatrix}$	(0,0,1,1,0)

2.1.1 Operation on Vectors

A few operations which are frequently needed as follows:

Addition

Two or more vectors of the same order can be added by correspondent components. That is,

$$\boxed{a+b=\begin{bmatrix}a_1+b_1\\a_2+b_2\\.\\.\\.\\a_n+b_n\end{bmatrix}}$$

Subtraction

The difference between two vectors a and b, of the same order, is defined to be that vector, a-b, which, when added to b, yields the vector a. Again, subtraction is performed componentwise. That is,

$$a-b=\begin{bmatrix} a_1-b_1 \\ a_2-b_2 \\ . \\ . \\ . \\ a_n-b_n \end{bmatrix}$$

Scalar Multiplication of a Vector

Assume we have some real number k -this is a scalar in vector algebra. Scalar multiplication of a vector involves multiplying each component of the vector by the scalar.

$$\boxed{ka=k\begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_n \end{bmatrix}=\begin{bmatrix} ka_1 \\ ka_2 \\ . \\ . \\ . \\ ka_n \end{bmatrix}}$$

Scalar Products of Two Vectors

Unlike scalar multiplications, in vector algebra, multiplication of two vectors need not lead to a vector. For example, one way of multiplying two vectors (of the same order of course) yields a *number* rather than a vector. This number is called their scalar product. To illustrate the scalar product of two vectors, consider the column vectors

$$a=\begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_n \end{bmatrix}, \quad \text{and} \quad b=\begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix}$$

Their scalar product is defined as

$$\boxed{\begin{aligned} &a'b=a_1b_1+a_2b_2+....+a_kb_k+...+a_nb_n \\ &=\sum_{k=1}^{n}a_kb_k \end{aligned}}$$

2.2 Matrices

A matrix A of order *m* by *n*, consists of a rectangular array of real numbers (scalars) arranged in *m* rows and *n* columns.

$$A_{3x3} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

It is often convenient to think of matrices as a collection of *n* columns of *m*-sized vectors.

2.2.1 Operations on Matrices

Matrix Addition

In matrix addition each entry of a sum matrix is the sum of the corresponding entries of the two matrices being added, again assuming they are of the same order. To illustrate,

$$A = \begin{bmatrix} -1 & 4 & 3 \\ 0 & 5 & 2 \end{bmatrix}; \quad B = \begin{bmatrix} 0 & 2 & 1 \\ -1 & 4 & 3 \end{bmatrix}; \quad A + B = C = \begin{bmatrix} -1 & 6 & 4 \\ -1 & 9 & 5 \end{bmatrix}$$

Matrix Multiplication

Matrices can also be multiplied by a number (scalar), and this is called scalar multiplication of the matrix. That is,

E = k**A**

if only if

$$(e_{ij}) = k(a_{ij})$$

for i = 1,2,… m; j = 1,2,…,n

2.2.2 Some Special Matrices

There are a number of special matrices that are very useful in statistical analysis. These include:

Diagonal, Scalar, Sign, and Identity Matrices

A special case of a symmetric matrix is a diagonal matrix. A diagonal matrix is defined as a square matrix in which all off-diagonal entries are zero. For example,

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad B = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}; \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$

If all entries on the main diagonal are equal scalars, then the diagonal matrix is called a *scalar matrix*. Examples are:

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}; B = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}; \quad C = \begin{bmatrix} -4 & 0 \\ 0 & -4 \end{bmatrix}$$

If some of the entries on the main diagonal are -1 and the rest are +1, the diagonal matrix is called a *sign matrix*. Examples are,

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad C = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

A *skew symmetric matrix* A is a square matrix in which all the diagonal elements a_{ij} are zero and A = -A'. For example,

$$\text{If} \quad A = \begin{bmatrix} 0 & -3 & 1 \\ 3 & 0 & 2 \\ 1 & -2 & 0 \end{bmatrix}; \text{then } -A' = \begin{bmatrix} 0 & -3 & -1 \\ 3 & 0 & 2 \\ -1 & -2 & 0 \end{bmatrix}$$

If the entries on the diagonal of a scalar matrix are each equal to unity, then this type of scalar matrix is called an *identity matrix,* denoted **I**. Examples are,

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

3.0 Application of Vector Space Operations to Statistical Data

Much of the foregoing discussion has been introduced for some specific purpose, namely, to describe matrix and vector operations that are relevant for statistical procedures. One of the main virtues of matrix algebra is its conciseness, that is, the succinct way which many statistical operations can be described.

To illustrate the compactness of vector space methods, consider the artificial data given below. The data was obtained from an experiment designed to obtain information about the willingness of water consumers to pay water rate given certain conditions about their situation.
The column Y shows the number of times a particular consumer has missed paying water rate in the last 20 years. The column marked X_1 shows the attitude rating of the consumer towards the water supply system and X_2 shows the number of years the consumer has been connected to the system. Even though this data is artificial, it represents a typical geospatial problem.

Consumer	**Y**	Y^2	X_1	X_1^2	X_2	X_2^2	YX_1	YX_2	X_1X_2
a	1	1	1	1	1	1	1	1	1

b	0	0	2	4	1	1	0	0	2
c	1	1	2	4	2	4	2	2	4
d	4	16	3	9	2	4	12	8	6
e	3	9	5	25	4	16	15	12	20
f	2	4	5	25	6	36	10	12	30
g	5	25	6	36	5	25	30	25	30
h	6	36	7	49	4	16	42	24	28
i	9	81	10	100	8	64	90	72	80
j	13	169	11	121	7	49	143	91	77
k	15	255	11	121	9	81	165	135	99
l	16	256	12	144	10	100	192	160	120
	75	***823***	***75***	***639***	***59***	***397***	***702***	***542***	***497***

Row cross-product matrix

$$B = \begin{array}{c|ccc} & Y & X_1 & X_2 \\ \hline Y & 823 & 702 & 542 \\ X_1 & 702 & 639 & 497 \\ X_2 & 542 & 497 & 397 \end{array}$$

SSCP matrix

$$S = \begin{array}{c|ccc} & Y & X_1 & X_2 \\ \hline Y & 354.25 & 233.25 & 173.25 \\ X_1 & 233.25 & 170.25 & 128.25 \\ X_2 & 173.25 & 128.25 & 106.92 \end{array}$$

Covariance matrix

$$C = \begin{array}{c|ccc} & Y & X_1 & X_2 \\ \hline Y & 29.52 & 19.44 & 14.44 \\ X_1 & 19.44 & 14.19 & 10.69 \\ X_2 & 14.44 & 10.69 & 8.91 \end{array}$$

Correlation matrix

$$R = \begin{array}{c|ccc} & Y & X_1 & X_2 \\ \hline Y & 1.00 & 0.95 & 0.89 \\ X_1 & 0.95 & 1.00 & 0.95 \\ X_2 & 0.89 & 0.95 & 1.00 \end{array}$$

Let us consider the role of matrix algebra in the development of data summaries to employing specific analytical techniques.

The computation of means, variances, covariances, correlations, etc., is a necessary preliminary to subsequent analyses in addition to being useful in its own right as a way to summarize aspects of variation in data.

3.1 Sums, Sums of Squares, and Cross Products

To demonstrate the compactness of matrix notation, suppose we are concerned with computing the usual sums, sums of squares, and sums of cross products of the "raw" scores involving, for example, Y and X_1 in the above table:

$$\sum Y; \sum X_1; \sum Y^2; \sum X_1^2; \sum YX_1$$

In scalar products form, the first two expressions are simply

$$\sum Y = 1'y = 75\,; \qquad \sum X_1 = 1'x_1 = 75$$

where $1'$ is a 1 X 12 unit vector, with all entries unity, and y and x_1 are the Y and X observation expressed as vectors. Notice in each case that a scalar product of two vectors is involved.

Similarly, the scalar product notion can be employed to compute three other quantities involving Y and X_1:

$$\sum Y^2 = y'y = 823 \qquad \sum X_1^2 = x_1'x_1 = 639 \qquad \sum YX_1 = y'x_1 = 702$$

The table above lists the numerical values for all of these products and, in addition, the products involving X_2 as well.

As a matter of fact, if we designated the matrix A to be the 12 X 3 matrix of original data involving variables Y, X_1, and X_2, the following expression

$$\boxed{B = A'A}$$

which is often called the *minor product moment* (of A), will yield a symmetric matrix B of order 3 X 3. The diagonal entries of the matrix B denote the raw sums of squares of each variable, and the off-diagonal elements denoted the raw sums of cross products as shown in the table. This shows the beauty of the matrix approach in the analysis.

3.2 Mean-Corrected (SSCP) Matrix

We can also express the sums of squares and cross products as deviations about the means of Y, X_1, and X_2. The mean-corrected sums of squares and cross-products matrix is often more simply called the SSCP (sums of squares and cross products) x notation as

$$\boxed{S = A'A - \frac{1}{m}(A'1)(1'A)}$$

Where 1 denotes a 12 X 1 unit vector and m denotes the number of observations; m = 12. The last term on the right-hand side of the equation represents the correction term and is a generalization of the usual scalar formula for computing sums of squares about the mean:
i.e.

$$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{m}$$

where $x = X - \overline{X}$; that is, where x denotes deviation-from-mean form. Alternatively, if the columnar means are subtracted out of A to begin with, yielding the mean-corrected matrix A_d, then

$$\boxed{S = A_d^{/} A_d}$$

For example, the mean-corrected sums of squares and cross products for Y and X_1 are

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{m} = 823 - \frac{(75)^2}{12} = 354.25$$

$$\sum x_1^2 = \sum X_1^2 - \frac{(\sum X_1)^2}{m} = 639 - \frac{(75)^2}{12} = 170.25$$

$$\sum yx_1 = \sum YX_1 - \frac{(\sum Y \sum X_1)}{m} = 702 - \frac{(75\ X\ 75)}{12} = 233.25$$

The SSC matrix for all three variables appears in the table above.

3.3 Covariance and Correlation Matrices

The covariance matrix, shown in the table, is obtained from the (mean-corrected) SSCP matrix by simply dividing each entry of S by the scalar m, the sample size. That is,

$$\boxed{C = \frac{1}{m} S}$$

In summational form the off-diagonal elements of C can be illustrated for the variables Y and X_1 by the notation

$$\operatorname{cov}(YX_1) = \Sigma yx_1 / m = 233.25 / 12 = 19.44$$

Note that a covariance, then, is merely an averaged cross product of mean-corrected scores. The diagonals of C are, of course, variances; for example,

$$sy^2 = \sum y^2 / m$$

The *correlation* between two variables, y and x_1, often obtained as

$$\boxed{r_{yx} = \frac{\sum yx_1}{\sqrt{\sum y^2 \sum x_1^2}}}$$

Where y and x_1 are each expressed in deviation-from-mean form (as noted above).

Not suprisingly, R the correlation matrix is related to S, the SSCP matrix, and C, the covariance matrix. For example, let us return to S. The entries on the main diagonal of S represent mean-corrected sums of squares of the three variables Y, X_1, and X_2.

If we take the square roots of these three entries and enter the reciprocals of these square roots in a diagonal matrix, we have

$$D = \begin{bmatrix} 1/\sqrt{\sum y^2} & 0 & 0 \\ 0 & 1/\sqrt{\sum x_1^2} & 0 \\ 0 & 0 & 1/\sqrt{\sum x_2^2} \end{bmatrix}$$

Then, by pre- and post multiplying S by D we can obtain the correlation matrix R.

$$\boxed{R = DSD}$$

$$R = \begin{bmatrix} \dfrac{\sum y}{\sum y^2\sqrt{\sum y^2}} & \dfrac{\sum yx_1}{\sqrt{\sum y^2}\sqrt{\sum x_1^2}} & \dfrac{\sum yx_2}{\sqrt{\sum y^2}\sqrt{\sum x_2^2}} \\ \dfrac{\sum yx_1}{\sqrt{\sum y^2}\sqrt{\sum x_1^2}} & \dfrac{\sum x_1^2}{\sqrt{\sum x_1^2}\sqrt{\sum x_1^2}} & \dfrac{\sum x_1x_2}{\sqrt{\sum x_1^2}\sqrt{\sum x_2^2}} \\ \dfrac{\sum yx_2}{\sqrt{\sum y^2}\sqrt{\sum x_2^2}} & \dfrac{\sum x_1x_2}{\sqrt{\sum x_1^2}\sqrt{\sum x_2^2}} & \dfrac{\sum x_2^2}{\sqrt{\sum x_2^2}\sqrt{\sum x_2^2}} \end{bmatrix}$$

The above matrix is the derived matrix of correlations between each pair of variables and is also shown in the table.

4.0 Summary

It is obvious from the foregoing that all the operations needed to compute the various cross-products used in analysis can be readily expressed in vector-space format. Specifically, matrix operations and vector representations in various forms can be used to formulate all the usual algorithms for statistical calculations. Using this approach has provided a rather compact and easy way to portray the calculations which otherwise are cumbersome to handle.

5.0 References

Green, P.E. and J.D. Carol (1976): *"Mathematical Tools for Applied Multivariate Analysis"*. Academic Press, New York.

Olaleye, J.B. (1992): *"Optimum Software Architecture for an Analytical Photogrammetric Workstation and its Integration into a Spatial Information Environment"*, Technical Report No. 162, Dept. of Surveying Engineering, University of New Brunswick, Canada, 228 pp.