Application of Linear Algebra to Statistics through the Mean Vector and Covariance Matrix

A presentation for the Directed Reading Program by Muhideen Ogunlowo

What is Linear Algebra and how can we apply it to statistics?

Linear algebra is the math of vectors and matrices.

In statistics, the main purpose of linear algebra is to organize data and write down the manipulations we want to do to them to get results for statistical analysis.

Matrices

A Matrix can be identified as an array of numbers or values usually arranged in rows or columns. Matrices come in M x N (Rectangular array) or N x N (Square array), where M is the number of rows and N is the number of columns. Example: A Matrix is said to be a 2 X 1 Matrix if it has two rows and 1 column.

Mean Vector and Covariance Matrix

Being the first step in analyzing multivariate data, the mean vector consists of the means of each variable and the variance-covariance matrix consists of the variances of the variables along the main diagonal and the covariances between each pair of variables in the other matrix positions. Basically, this works like your everyday mean and variance except there is a little twist to it. It comes in a Matrix form.

The formula for calculating the covariance of any two variables is

$$\mathrm{COV} = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})}{n-1},$$

where $\bar{x}$ is the mean vector for a set of variables and $\bar{y}$ is the mean vector of another set of variables and n is the number of values or data corresponding to a variable.

Sample Data Matrix

Consider the following Matrix A where each column represents a different variable. Each row vector Ai is an observation of the three variables assumed to be the length, breadth and width of different models of a solar panel for a developing community.

4.1	2.1	0.61
4.6	2.0	0.59
3.9	1.8	0.57
4.5	1.7	0.58
4.7	1.9	0.63

The result of mean vector which contains the statistical averages of each variable of the Matrix is

4.36	1.92	0.596

Therefore, after an application of the formula given, the result of the covariance matrix is

0.1180	0.0060	0.0041
0.0060	0.0270	0.0029
0.0041	0.0029	0.0006

Using the results gained from this experiment, we can now deduce that 0.1180 is the variance of the length variables, 0.0060 is the covariance between the length and breadth variables, and 0.0041 is the covariance between the length and height variables. The variance of the breadth variables is 0.0270 and 0.0006 is the variance of the height variables.

You may notice that the diagonal values of the matrix refers to the variances of each variable with respect to itself and other values of the covariance matrix refer to variances of each variable with respect to another variable.

Another key point is that Covariance Matrices always come in a square matrix format where the rows and columns are of an equal dimension. The mean vector is always in a 1 X n matrix, where n is the number of variables or columns of your data set.

The mean vector is often referred to as the centroid and the

Practical applications of the Mean Vector and Covariance Matrix

Covariance measures how much two random variables vary together in a population. When the population contains higher dimensions or more random variables, a matrix is used to describe the relationship between different dimensions. To make things simple, covariance matrix is to define the relationship in the entire dimensions as the relationships between every two random variables.

Covariance is a significant tool in modern portfolio theory used to ascertain what securities to put in a portfolio.