Centroid

Hervé Abdi*

The concept of centroid is the multivariate equivalent of the mean. Just like the mean, the centroid of a cloud of points minimizes the sum of the squared distances from the points of the cloud to a point in the space . © 2009 John Wiley & Sons, Inc. Centroid

The notion of centroid generalizes the notion of a mean to multivariate analysis and multidimensional spaces. It applies to vectors instead of scalars, and it is computed by associating to each vector a mass which is a positive number taking values between 0 and 1 and such that the sum of all the masses is equal to 1. The centroid of a set of vectors is also called the *center of gravity*, the *center of mass*, or the *barycenter* of this set.

NOTATIONS AND DEFINITION

Let $\mathcal{V}$ be a set of I vectors with each vector being composed of J elements

$$\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_I\} \text{ with}$$
$$\mathbf{v}_i = [v_{i,1}, \ldots, v_{i,j}, \ldots, v_{i,J}]^\mathsf{T}. \quad (1)$$

To each vector is associated a mass denoted m_i for vector i. These masses take values between 0 and 1 and the sum of these masses is equal to 1. The set of masses is a vector denoted $\mathbf{m}$. The centroid of the set of vectors is denoted $\mathbf{c}$, it is defined as

$$\mathbf{c} = \sum_i^I m_i \mathbf{v}_i. \quad (2)$$

EXAMPLES

The mean of a set of numbers is the centroid of this set of observations. Here, the mass of each number is equal to the inverse of the number of observations: $m_i = \frac{1}{I}$.

---

*Correspondence to: herve@utdallas.edu

The University of Texas at Dallas, Richardson, TX, USA

For multivariate data, the notion of centroid generalizes the mean. For example, with the following three vectors:

$$\mathbf{v}_1 = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 8 \\ 24 \end{bmatrix}, \text{ and } \mathbf{v}_3 = \begin{bmatrix} 16 \\ 32 \end{bmatrix}, \quad (3)$$

and the following set of masses:

$$m_1 = \frac{1}{2}, \quad m_2 = \frac{1}{8}, \quad \text{and} \quad m_3 = \frac{3}{8}, \quad (4)$$

we obtain the following centroid

$$\mathbf{c} = \sum_i^I m_i \mathbf{v}_i = \frac{1}{2} \begin{bmatrix} 10 \\ 20 \end{bmatrix} + \frac{1}{8} \begin{bmatrix} 8 \\ 24 \end{bmatrix} + \frac{3}{8} \begin{bmatrix} 16 \\ 32 \end{bmatrix} = \begin{bmatrix} 12 \\ 25 \end{bmatrix}. \quad (5)$$

In this example, if we plot the vectors in a two-dimensional space, the centroid would be the center of gravity of the triangle made by these three vectors with masses assigned proportionally to their vector of mass. The notion of centroid can be used with spaces of any dimensionality.

PROPERTIES OF THE CENTROID

The properties of the centroid of a set of vectors closely parallel the more familiar properties of the mean of a set of numbers. Recall that a set of vectors defines a multidimensional space, and that to each multidimensional space is assigned a generalized Euclidean distance. The core property of the centroid is that: the centroid of a set of vectors minimizes the weighted sum of the generalized squared Euclidean distances from the vectors to any point in the space. This quantity which generalizes the notion of variance, is called the *Inertia* of the set of vectors relative to their centroid.

Of additional interest for multivariate analysis, the *theorem of Huyghen's* indicates that the weighted sum of the squared distances from a set of vectors to any vector in the space can be decomposed as a weighted sum of the squared distances from the vectors to their centroid plus the (weighthed) squared distance from the centroid to this point. In term of inertia, Huyghen's theorem states that the inertia of a set of vectors to any point is equal to the inertia of the set of vectors to their centroid plus the inertia of their centroid to this point. As an obvious consequence of this theorem, the inertia of a set of vectors to their centroid is minimal. Huyghen's theorem is the basis of several statistical methods such as Analysis of Variance and Discriminant Analysis. In these techniques, the total Inertia of a cloud of points is partitioned into the Inertia of the points to their group centroid (i.e., the within group Inertia) plus the Inertia of the group centroids to their own centroid.