Data Matrices and Linear Algebra

The Geometry of Statistics

Ramesh Srinivasan

October 24, 2023

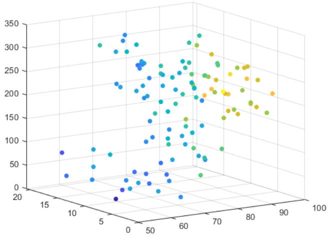
Multivariate Data Structure

A data matrix of the form below is at the heart of any data science project.

$$Data_k = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ \vdots & \ddots & & \\ a_{N1} & \dots & & a_{NM} \end{bmatrix}$$

- In this data matrix there are M columns corresponding to the M different variables being measured.
- In this data matrix there are N rows corresponding to N observations
- If there are multiple data collection, there may be K such matrices.
- The goal is to take the data in this matrix and:
  1. *Classification/Regression* Use the data to predict another variable. In psychology, this is usually behavior.
  2. *Clustering* Use the data to learn about supgroups of the data.
  3. *Latent Variable Models* Learn about hidden variables that are generating the observed variables.

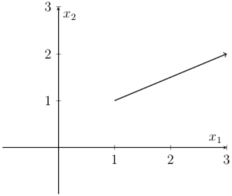
Multivariate Embedding



I am going to represent the data as points (actually vectors) in a multidimensional space.
Ideas from Linear Algebra will inform us on how to think about the data.

Vectors

A vector is a combination of numbers representing a magnitude and a direction.
They are defined by an *origin* and an *endpoint*.
For example, in the 2D plane, (x_1, x_2) we can have a vector originating at coordinate (1,1) and ending at (3,2)



N-Dimensional Vectors

In the 2D vector example above, the coordinate of the origin and endpoint has a component x_1 and a component x_2 .

A vector can be defined in **any** number of dimensions,
e.g., a vector can have origin (0,0,0,0) and endpoint (1,2,3,4) in 4 dimensions.

In the most general case, a vector can be defined in n-dimensional space by the coordinate of its origin, and endpt which will have components $(x_1, x_2, ..., x_n)$

A vector can be translated to have origin at $\mathbf{0} = (0, 0, 0, 0.....0)$ by subtracting the origin from the endpoint in each dimension.

When a vector is specified with only one coordinate it is often implicit that the origin of the vector is at the origin of the coordinate system.

Vector Norm

The length of a vector is a type of vector norm or measure of magnitude of a vector. (Specifically, its the L2 norm)

In two dimensions, if $\mathbf{x} = (x_1, x_2)$

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}$$

In n dimensions, if $\mathbf{x} = (x_1, x_2, x_3, ...x_n)$

$$\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2}$$

Unit Vectors

A vector of length 1, i.e., $\|\mathbf{u}\| = 1$ is known as a unit vector.
A unit vector $\mathbf{u}$ has the same direction as the vector $\mathbf{x}$ if

$$\mathbf{u} = \frac{\mathbf{X}}{\|\mathbf{x}\|}$$

e.g., $\mathbf{u}$ = (3/5,4/5) is a unit vector in the same direction as $\mathbf{x}$ = (3,4)

The coordinate axes are n-dimensional unit vectors,
$\mathbf{u}_1 = (1, 0, ...0)$
$\mathbf{u}_2 = (0, 1, ...0)...$
$\mathbf{u}_n = (0, 0, ...1)$

Dot product or Inner Product

in 2-D,

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2$$

in 3-D,

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + x_3 y_3$$

in n-D,

$$\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k$$

The dot product of a vector with itself its norm(length) squared

$$\mathbf{x} \cdot \mathbf{x} = \sum_{k=1}^{n} x_k x_k = \|x\|^2$$

Meaning of dot product

The dot product has a physical interpretation. The dot product is proportional to the cosine of the angle between two vectors.

$$\mathbf{x} \cdot \mathbf{y} = \|x\| \, \|y\| \, cos(\theta)$$

If two vectors are parallel, the dot product is the product of their lengths. If the two vectors are perpendicular the dot product is zero.

Orthogonal and Orthonormal vectors

Two vectors are orthogonal if their dot product is zero,

$$\mathbf{x} \cdot \mathbf{y} = 0$$

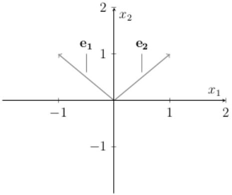
Two vectors are *orthonormal* if their dot product is zero and they have length 1:

$$\|x\| = \|y\| = 1$$

$$\mathbf{x} \cdot \mathbf{y} = 0$$

Basis of a Vector Space

In 2-D, unit vectors along the coordinate axes $\mathbf{u_1} = (1, 0)$ and $\mathbf{u_2} = (0, 1)$ are orthonormal vectors.
Any vector $\mathbf{x}$ in the plane can be written as a linear combination, $\mathbf{x} = x_1\mathbf{u_1} + x_2\mathbf{u_2}$
Thus, together $\{\mathbf{u_1}, \mathbf{u_2}\}$ *span* the vector space of all vectors in a plane, and form a **basis** of the vector space.



In 2-dimensional space, any 2 linearly independent vectors can form an basis.
If n-dimensional space, any n linearly independent vectors can form a basis.
Linearly independent vectors have dot product of zero.

Geometry of Multivariate Data

n observations of p variables can be represented as a $n \times p$ matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \ddots & & \\ x_{n1} & \dots & & x_{np} \end{bmatrix}$$

n observations may represent n different participants in an experiment while p are different experimental variables observed in each participant.

In Neuroscience applications, n represents different samples in time, while p represents different locations in the brain.

$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_n} \end{bmatrix}$ The matrix a stack of n row vectors (of size p) of observations. **This is how data is collected.**

$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} \dots \mathbf{x_p} \end{bmatrix}$ The matrix is a stack of p column vectors (of size n) of variables. **This is the useful way to think about data**

Centering Data

$\bar{\mathbf{X}}$ is a row vector of length p which is the coordinates given by averaging each column of $\mathbf{X}$.
The components of $\bar{\mathbf{X}}$ are

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, k = 1, 2, ...p$$

When performing data analysis, we should always center the data on the origin of the coordinate system by computing the *deviations*

$$\mathbf{d_k} = \mathbf{x_k} - \bar{x}_k, k = 1, 2, ...p$$

Example

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

$$\bar{\mathbf{X}} = [2 \quad 3]$$

$$\mathbf{d}_1 = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\mathbf{d}_2 = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 2 & -2 \\ -3 & 0 \\ 1 & 2 \end{bmatrix}$$

D is a matrix of deviations, which is the original data matrix X now centered on the origin of the coordinate system.

Standard Deviation is a measure of length or norm

If we compute the squared length or norm of a deviation vector, we get a measure of variance,

$$\|\mathbf{d_k}\|^2 = \mathbf{d_k} \cdot \mathbf{d_k} = \sum_{j=1}^{n} d_{jk}^2 = \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 = n s_k^2$$

Therefore the length of the vector is proportional to standard deviation.

$$s_k = \sqrt{\frac{1}{n} \mathbf{d_k} \cdot \mathbf{d_k}} = \sqrt{\frac{1}{n} \|\mathbf{d_k}\|^2}$$

Covariance and Correlation Coefficient

Covariance is related to the dot product between two different deviation vectors

$$s_{kl} = \frac{1}{n}\mathbf{d_k} \cdot \mathbf{d_l}$$

Correlation coefficient is the dot product of unit vectors in the direction of the two data vectors.

$$r_{kl} = \frac{s_{kl}}{s_k s_l} = \frac{\mathbf{d_k} \cdot \mathbf{d_l}}{\|\mathbf{d_k}\| \, \|\mathbf{d_l}\|} = \mathbf{u_k} \cdot \mathbf{u_l}$$