Multivariate Analysis

Course: STAT-403

Prof. Dr. Rumana Rois
Department of Statistics and Data Science
Jahangirnagar University

What is multivariate analysis?

The iris data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant, for instance, Setosa, Versicolour, and Virginica.

Table 1: IRIS DATA.

Sepal Length in cm (x_1)	Sepal Width in cm (x_2)	Petal Length in cm (x_3)	Petal Width in cm (x_4)	Class (x_5)
5.1	3.5	1.4	0.2	1
4.9	3.0	1.4	0.2	1
4.7	3.2	1.3	0.2	1
⋮	⋮	⋮	⋮	⋮
6.2	3.4	5.4	2.3	3
5.9	3.0	5.1	1.8	3

What is multivariate analysis?

The iris data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant, for instance, Setosa, Versicolour, and Virginica.

Table 1: IRIS DATA.

Sepal Length in cm (x_1)	Sepal Width in cm (x_2)	Petal Length in cm (x_3)	Petal Width in cm (x_4)	Class (x_5)
5.1	3.5	1.4	0.2	1
4.9	3.0	1.4	0.2	1
4.7	3.2	1.3	0.2	1
⋮	⋮	⋮	⋮	⋮
6.2	3.4	5.4	2.3	3
5.9	3.0	5.1	1.8	3

Multivariate analysis is concerned with the data that consist of simultaneous measurements on many variables.

Multivariate analysis

Univariate vs. Multivariate

1. Univariate analysis is used when one variable is measured for each observation.

Multivariate analysis

Univariate vs. Multivariate

1. Univariate analysis is used when one variable is measured for each observation.
2. Multivariate analysis is used when more than one outcome variables are measured for each observation.

Multivariate analysis

Univariate vs. Multivariate

1. Univariate analysis is used when one variable is measured for each observation.
2. Multivariate analysis is used when more than one outcome variables are measured for each observation.

Dependence and Interdependence techniques

Multivariate analysis

Univariate vs. Multivariate

1. Univariate analysis is used when one variable is measured for each observation.
2. Multivariate analysis is used when more than one outcome variables are measured for each observation.

Dependence and Interdependence techniques

1. Dependence techniques are appropriate when one or more variables can be identified as dependent variables and the remaining as independent variables.

Multivariate analysis

Univariate vs. Multivariate

1. Univariate analysis is used when one variable is measured for each observation.
2. Multivariate analysis is used when more than one outcome variables are measured for each observation.

Dependence and Interdependence techniques

1. Dependence techniques are appropriate when one or more variables can be identified as dependent variables and the remaining as independent variables.
2. In interdependence techniques, the variables are not classified as dependent or independent, rather the whole set of independent relations is examined.

Multivariate analysis

Dependence and Interdependence techniques

1. **Dependence techniques** are appropriate when one or more variables can be identified as dependent variables and the remaining as independent variables.
2. In **interdependence techniques**, the variables are not classified as dependent or independent, rather the whole set of independent relations is examined.

Multivariate analysis

Dependence and Interdependence techniques

1. **Dependence techniques** are appropriate when one or more variables can be identified as dependent variables and the remaining as independent variables.
    1. Multivariate Analysis Of Variance (MANOVA) and Covariance (MANCOVA), Multiple Discrimination Analysis, Multivariate Regression.
2. In **interdependence techniques**, the variables are not classified as dependent or independent, rather the whole set of independent relations is examined.

Multivariate analysis

Dependence and Interdependence techniques

1. **Dependence techniques** are appropriate when one or more variables can be identified as dependent variables and the remaining as independent variables.
    1. Multivariate Analysis Of Variance (MANOVA) and Covariance (MANCOVA), Multiple Discrimination Analysis, Multivariate Regression.
2. In **interdependence techniques**, the variables are not classified as dependent or independent, rather the whole set of independent relations is examined.
    1. Factor Analysis, Cluster Analysis, Canonical correlation, Principal Components Analysis, Multidimensional Scaling.

Applications of multivariate analysis

1. Data reduction or structural simplification
2. Sorting and grouping
3. Investigation of the dependence among variables
4. Prediction
5. Hypothesis construction and testing

Organization of Data

x_{jk}= measurements of the kth variable on the jth item

Organization of Data

x_{jk}= measurements of the kth variable on the jth item

n measurements on p variables can be displayed as

	Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Matrix Algebra and Random Vectors

2.5 RANDOM VECTORS AND MATRICES

A *random vector* is a vector whose elements are random variables. Similarly, a *random matrix* is a matrix whose elements are random variables. The expected value of a random matrix (or vector) is the matrix (vector) consisting of the expected values of each of its elements. Specifically, let $\mathbf{X} = \{X_{ij}\}$ be an $n \times p$ random matrix. Then the expected value of $\mathbf{X}$, denoted by $E(\mathbf{X})$, is the $n \times p$ matrix of numbers (if they exist)

$$E(\mathbf{X}) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & \cdots & E(X_{1p}) \\ E(X_{21}) & E(X_{22}) & \cdots & E(X_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n1}) & E(X_{n2}) & \cdots & E(X_{np}) \end{bmatrix} \quad (2\text{-}23)$$

where, for each element of the matrix,[2]

$$E(X_{ij}) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x_{ij} f_{ij}(x_{ij})\, dx_{ij} & \text{if } X_{ij} \text{ is a continuous random variable with probability density function } f_{ij}(x_{ij}) \\ \displaystyle\sum_{\text{all } x_{ij}} x_{ij} p_{ij}(x_{ij}) & \text{if } X_{ij} \text{ is a discrete random variable with probability function } p_{ij}(x_{ij}) \end{cases}$$

Mean Vectors and Covariance Matrices

$$\mu_i = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with probability density function } f_i(x_i) \\ \displaystyle\sum_{\text{all } x_i} x_i p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases}$$

$$\sigma_i^2 = \begin{cases} \displaystyle\int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with probability density function } f_i(x_i) \\ \displaystyle\sum_{\text{all } x_i} (x_i - \mu_i)^2 p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with probability function } p_i(x_i) \end{cases} \quad (2\text{-}25)$$

Mean Vectors and Covariance Matrices

$$
\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k)
$$

$$
= \begin{cases} \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k)\,dx_i\,dx_k & \text{if } X_i, X_k \text{ are continuous random variables with the joint density function } f_{ik}(x_i, x_k) \\ \displaystyle\sum_{\text{all } x_i}\sum_{\text{all } x_k} (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) & \text{if } X_i, X_k \text{ are discrete random variable with joint probability function } p_{ik}(x_i, x_k) \end{cases} \tag{2-26}
$$

Mean Vectors and Covariance Matrices

$$
\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'
$$

$$
= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} [X_1 - \mu_1, X_2 - \mu_2, \ldots, X_p - \mu_p] \right)
$$

$$
= E \begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & (X_p - \mu_p)(X_2 - \mu_2) & \cdots & (X_p - \mu_p)^2 \end{bmatrix}
$$

$$
= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \cdots & E(X_1 - \mu_1)(X_p - \mu_p) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 & \cdots & E(X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p - \mu_p)(X_1 - \mu_1) & E(X_p - \mu_p)(X_2 - \mu_2) & \cdots & E(X_p - \mu_p)^2 \end{bmatrix}
$$

Find the mean and covariance matrices

Consider the random vector $X' = [X_1, X_2]$. Let the discrete random variable X_1 have the probability function p_1, X_2 have p_2 and their joint probability function $p_{12}(x_1, x_2)$.

x_1 \ x_2	0	1	$p_1(x_1)$
−1	.24	.06	.3
0	.16	.14	.3
1	.40	.00	.4
$p_2(x_2)$	.8	.2	1

Population Correlation matrix

Let the population correlation matrix be the $p \times p$ symmetric matrix

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \qquad (2\text{-}34)$$

and let the $p \times p$ *standard deviation* matrix be

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix} \qquad (2\text{-}35)$$

Then it is easily verified (see Exercise 2.23) that

$$\mathbf{V}^{1/2}\boldsymbol{\rho}\mathbf{V}^{1/2} = \boldsymbol{\Sigma} \qquad (2\text{-}36)$$

Find the mean, covariance and correlation matrices

Consider the random vector $X' = [X_1, X_2]$. Let the discrete random variable X_1 have the probability function p_1, X_2 have p_2 and their joint probability function $p_{12}(x_1, x_2)$.

x_1 \ x_2	0	1	$p_1(x_1)$
−1	.24	.06	.3
0	.16	.14	.3
1	.40	.00	.4
$p_2(x_2)$	.8	.2	1

Mean and Covariance of Linear Combinations of Matrices

2.30. You are given the random vector $\mathbf{X}' = [X_1, X_2, X_3, X_4]$ with mean vector $\boldsymbol{\mu}'_{\mathbf{X}} = [4, 3, 2, 1]$ and variance–covariance matrix

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 9 & -2 \\ 2 & 0 & -2 & 4 \end{bmatrix}$$

Partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \hdashline X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \hdashline \mathbf{X}^{(2)} \end{bmatrix}$$

Let

$$\mathbf{A} = [1 \quad 2] \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 1 & -2 \\ 2 & -1 \end{bmatrix}$$

and consider the linear combinations $\mathbf{AX}^{(1)}$ and $\mathbf{BX}^{(2)}$. Find

(a) $E(\mathbf{X}^{(1)})$

(b) $E(\mathbf{AX}^{(1)})$

(c) $\text{Cov}(\mathbf{X}^{(1)})$

(d) $\text{Cov}(\mathbf{AX}^{(1)})$

(e) $E(\mathbf{X}^{(2)})$

(f) $E(\mathbf{BX}^{(2)})$

Distance

Consider the point $P = (x_1, x_2)$ in the plane. The straight line (Euclidian) distance, $d(O, P)$, from P to the origin $O = (0, 0)$ is (Pythagoras)

$$d(O, P) = \sqrt{x_1^2 + x_2^2}.$$

In general, if P has p coordinates so that $P = (x_1, x_2, \ldots, x_p)$, the Euclidian distance is

$$d(O, P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}.$$

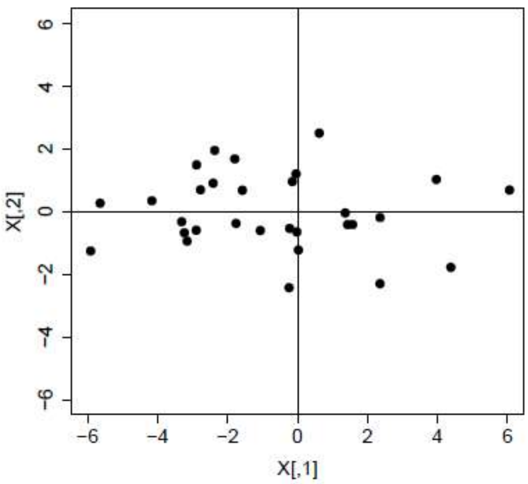
The distance between 2 arbitrary points P and $Q = (y_1, y_2, \ldots, y_p)$ is given by

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}.$$

Each coordinate contributes equally to the calculation of the Euclidian distance. It is often desirable to weight the coordinates.

Statistical Distance

Statistical distance should account for differences in variation and correlation. Suppose we have n pairs of measurements on 2 independent variables x_1 and x_2:



Variability in x_1 direction is much larger than in x_2 direction! Values that are a given deviation from the origin in the x_1 direction are not as *surprising* as are values in x_2 direction.
It seems reasonable to weight an x_2 coordinate more heavily than an x_1 coordinate of the same value when computing the distance to the origin.

Distance

Compute the statistical distance from the standardized coordinates

$$x_1^* = \frac{x_1}{\sqrt{s_{11}}} \quad \text{and} \quad x_2^* = \frac{x_2}{\sqrt{s_{22}}}$$

as

$$d(O,P) = \sqrt{(x_1^*)^2 + (x_2^*)^2} = \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{s_{22}}}\right)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}.$$

This can be generalized to accommodate the calculation of statistical distance from an arbitrary point $P = (x_1, x_2)$ to any *fixed* point $Q = (y_1, y_2)$. If the coordinate variables vary independent of one other, the distance from P to Q is

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}}.$$

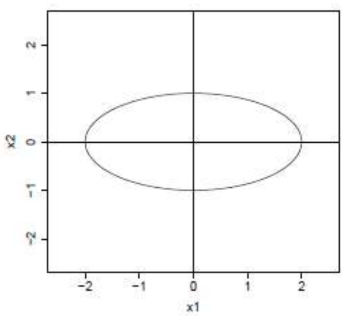
The extension to more than 2 dimensions is straightforward.

Distance

Let $P = (x_1, x_2, \ldots, x_p)$ and $Q = (y_1, y_2, \ldots, y_p)$. Assume again that Q is fixed. The statistical distance from P to Q is

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}.$$

- The distance of P to the origin is obtained by setting $y_1 = y_2 = \cdots = y_p = 0$.
- If $s_{11} = s_{22} = \cdots = s_{pp}$, the Euclidian distance is appropriate.



Consider a set of paired measurements (x_1, x_2) with $\overline{x}_1 = \overline{x}_2 = 0$, and $s_{11} = 4$, $s_{22} = 1$. Suppose the x_1 measurements are unrelated to the x_2 ones. We measure the squared distance of an arbitrary $P = (x_1, x_2)$ to $(0,0)$ by $d^2(O,P) = x_1^2/4 + x_2^2/1$. All points with constant distance 1 satisfy: $x_1^4/4 + x_2^2/1 = 1$, an Ellipse centered at $(0,0)$.

Properties of Distance

1. $d(P, Q) = d(Q, P)$,
2. $d(P, Q) > 0$ if $P \neq Q$,
3. $d(P, Q) = 0$ if $P = Q$,
4. $d(P, Q) \leq d(P, R) + d(R, Q)$, R being any other point different to P and Q.

Ellipse of constant Statistical Distance

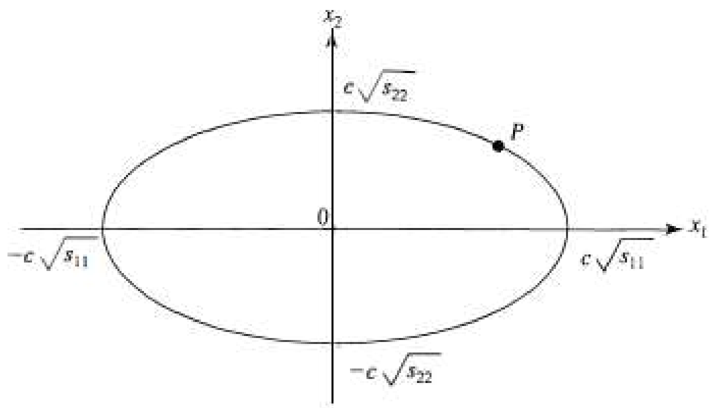


Figure 1.21 The ellipse of constant statistical distance $d^2(O, P) = x_1^2/s_{11} + x_2^2/s_{22} = c^2$.

Quadratic Form of Distance

Definition 2A.32. A *quadratic form* $Q(\mathbf{x})$ in the k variables $x_1, x_2, \ldots, x_k$ is $Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{x}' = [x_1, x_2, \ldots, x_k]$ and $\mathbf{A}$ is a $k \times k$ symmetric matrix.

Note that a quadratic form can be written as $Q(\mathbf{x}) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} x_i x_j$. For example,

$$Q(\mathbf{x}) = [x_1 \quad x_2] \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 2x_1x_2 + x_2^2$$

$$Q(\mathbf{x}) = [x_1 \quad x_2 \quad x_3] \begin{bmatrix} 1 & 3 & 0 \\ 3 & -1 & -2 \\ 0 & -2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 + 6x_1x_2 - x_2^2 - 4x_2x_3 + 2x_3^2$$

Any symmetric square matrix can be reconstructured from its eigenvalues and eigenvectors. The particular expression reveals the relative importance of each pair according to the relative size of the eigenvalue and the direction of the eigenvector.

Spectral Decomposition of a Matrix

Let $\mathbf{A}$ be a $k \times k$ positive definite matrix with the spectral decomposition $\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i'$. Let the normalized eigenvectors be the columns of another matrix $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k]$. Then

$$\underset{(k\times k)}{\mathbf{A}} = \sum_{i=1}^{k} \lambda_i \underset{(k\times 1)}{\mathbf{e}_i} \underset{(1\times k)}{\mathbf{e}_i'} = \underset{(k\times k)}{\mathbf{P}} \underset{(k\times k)}{\mathbf{\Lambda}} \underset{(k\times k)}{\mathbf{P}'} \qquad (2\text{-}20)$$

where $\mathbf{PP}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$ and $\mathbf{\Lambda}$ is the diagonal matrix

$$\underset{(k\times k)}{\mathbf{\Lambda}} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \quad \text{with } \lambda_i > 0$$