Analiza wariancji i metody klasyfikacyjne

Analysis of variance and classification methods

7 October 2019

Ilona Anna Urbaniak (PK)

Marcin Wolter (IFJ PAN)

e-mail: marcin.wolter@ifj.edu.pl, phone: 12 662 8024

Slides: https://indico.ifj.edu.pl/event/271/

ANalysis Of Variance - ANOVA

- Analysis of variance (ANOVA) - a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample.
- Based on the *“law of total variance”* - the observed variance in a particular variable is partitioned into components attributable to different sources of variation.
- Simplest form - a statistical test whether two or more population means are equal, and therefore generalizes the Student’s t-test beyond two means.
- ANOVA was developed by sir Ronald Fisher in 1920’ies (statistician and evolutionary biologist).

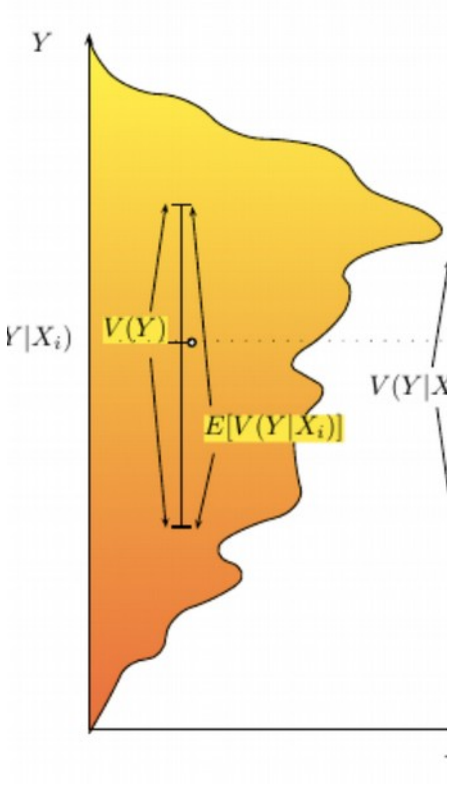
Some history

- Laplace was performing hypothesis testing already in the 1770s
- Least-squares methods developed by Laplace and Gauss circa 1800
- By 1827, Laplace was using least squares methods to address ANOVA problems regarding measurements of atmospheric tides.

- Sir Ronald Fisher introduced the term “variance” and proposed its formal analysis in a 1918 article *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*.
- His first application of the analysis of variance was published in 1921 *On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample*. Analysis of variance became widely known after being included in his 1925 book *Statistical Methods for Research Workers*.
- Important works by Jerzy Spława Neyman. First in 1923 *“On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.”*

Simple example

- Distribution of weights of dogs presented on a dog show.
- Task: predict the weight of a dog based on a certain set of characteristics of each dog (old-young, long-short hair, breed etc)
- Solution: “explain” the distribution of weights by dividing the dog population into groups based on those characteristics.
- Split dogs such that
  - each group has a low variance of dog weights (meaning the group is relatively homogeneous)
  - the mean of each group is distinct (if two groups have the same mean, then it isn't reasonable to conclude that the groups are, in fact, separate in any meaningful way).



Simple example

- Split by old-young and long-short hair
- The distributions have huge variance and similar means: not a good choice!
- Try something else...

Young, short-haired

Old, short-haired

Young, long-haired

Old, long-haired

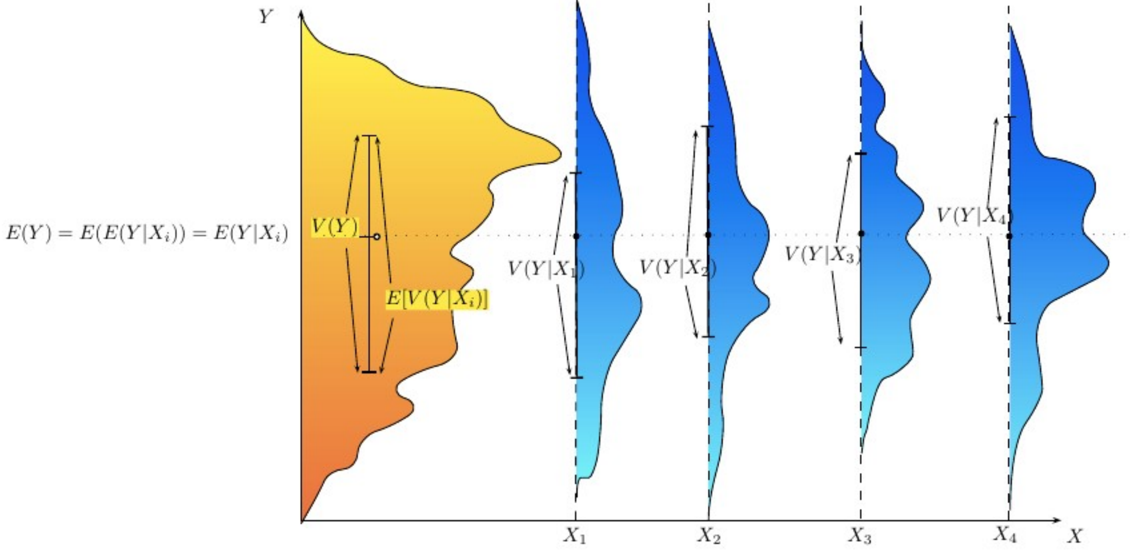


Figure 2: ANOVA : No fit

Simple example

- Now the dogs are split according to: pet vs working breed and less athletic vs more athletic.
- Partially successful, smaller variance and different means.
- We cannot distinguish X_1 and X_2 reliably.

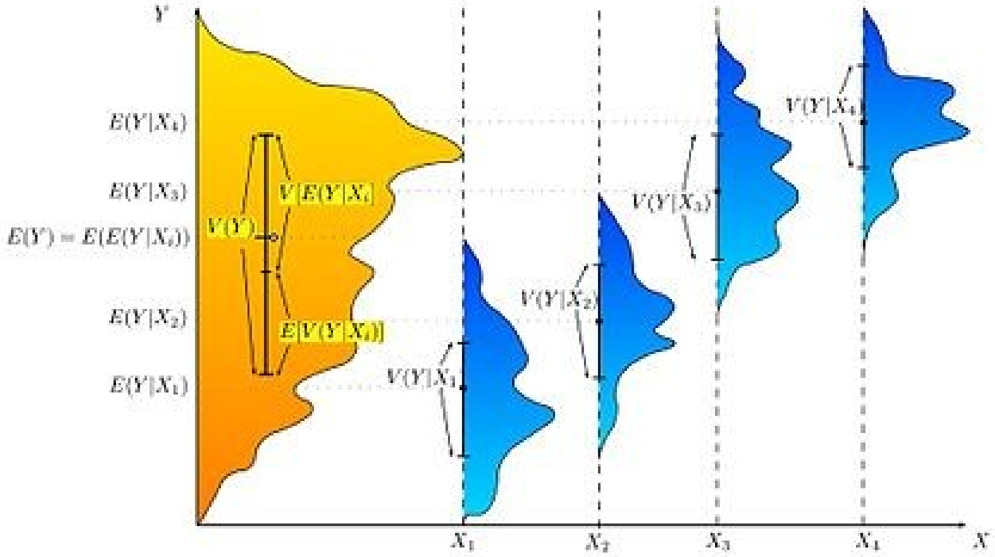


Figure 1: ANOVA : Fair fit

pet vs working breed and less athletic vs more athletic

Simple example

- Explain the dog weight by breed – much more successful
- Small variances, distinct means.

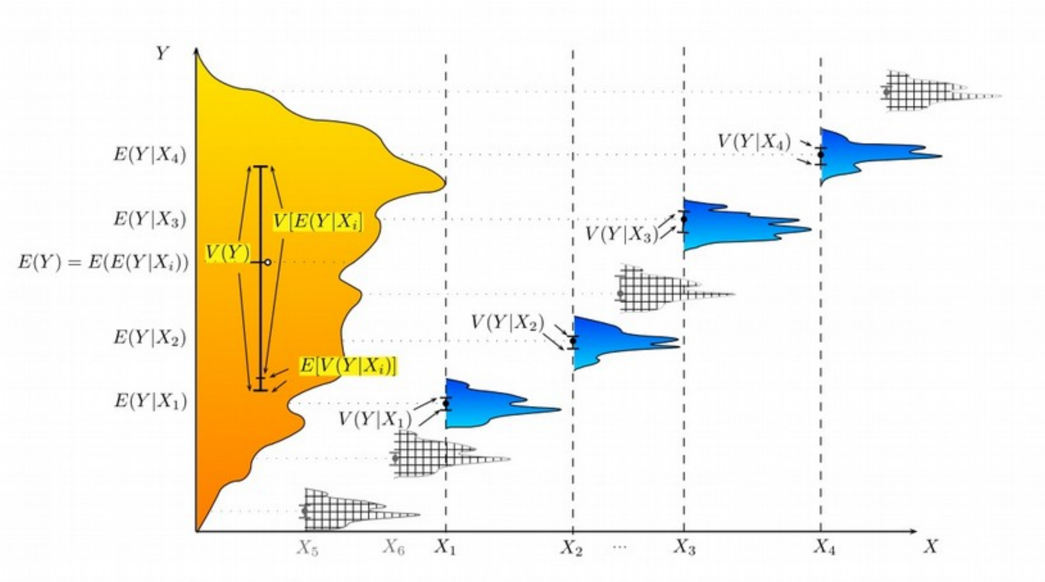


Figure 3: ANOVA : very good fit

ANOVA provides the formal tools to do, what was done intuitively here. It is used for analysis of experimental data.

Outline

W1	Przypomnienie podstawowych pojęć z teorii macierzy. Wielowymiarowy rozkład normalny. Macierz wariancji i kowariancji. Formy kwadratowe oparte o rozkład normalny.	4
W2	Wnioskowanie dotyczące dwóch średnich. Test t Studenta. Wnioskowanie dotyczące wielu średnich. Test F Fishera Snedecora.	4
W3	Jedno i dwukierunkowa analiza wariancji. Test interakcji.	4
W4	Metoda składowych głównych (PCA). Wybór optymalnego układu współrzędnych.	4
	Redukcja wymiarowości danych.	
W5	Klasyfikacja danych jedno i wielowymiarowych. Klasyfikator Fisher'a i inne proste klasyfikatory.	4
W6	Analiza skupisk.	4
W7	Wstęp do uczenia maszynowego	4
W8	Powtórka materiału	2

My first lecture on PK, I learned about it 10 days ago. The program might change :)

Bibliography

[1] Johnson, Wichern — Applied *Multivariate Statistical Analysis*

http://docshare04.docshare.tips/files/12598/125983744.pdf

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani -- *An Introduction to Statistical Learning*

http://faculty.marshall.usc.edu/gareth-james/ISL/

[3] John A. Rice -- *Mathematical Statistics and Data Analysis*

https://epdf.pub/mathematical-statistics-and-data-analysis65096.html

And maybe something more...

Repetition:
linear algebra, multivariate distributions

- Random vectors
- Independence
- Expectations and Covariances
- IQuadratic Forms
- Multivariate Normal Distribution
- Why normal distributions are so important? Central Limit Theorem (CLT)

Random vector

- Definition

We define X_p to be a p-variate random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

where its entries X_1 , . . . , X_p are random variables .

- Remark

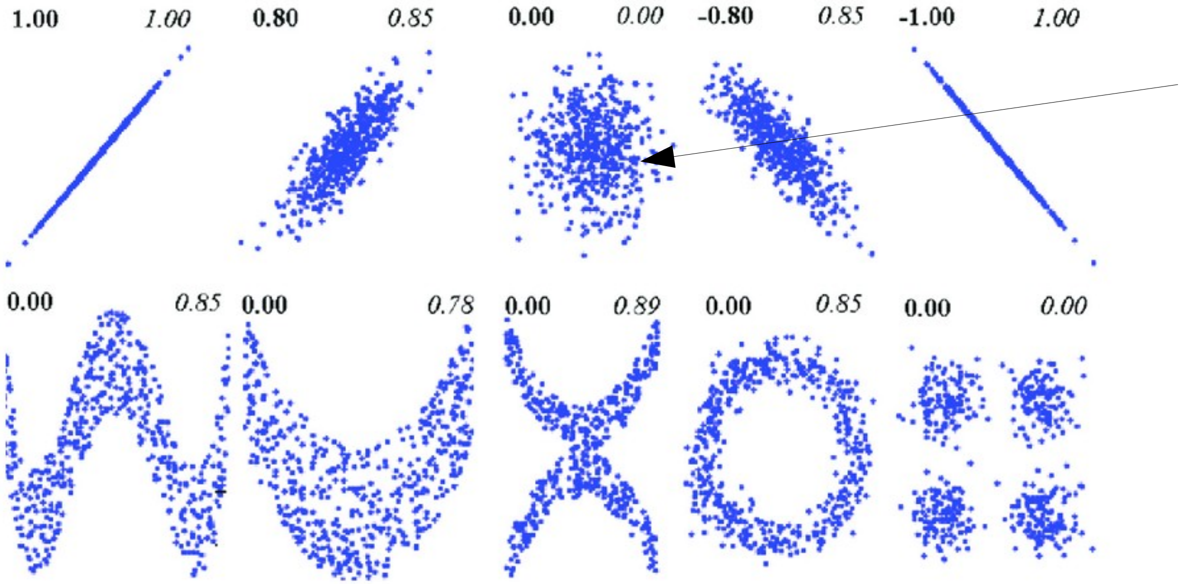
A random variable can be considered a univariate random vector.

Independence

If $X_1 , . . . , X_m$ are continuous, independent implies that the density factorizes:

$$\mathrm{f}_{1,\dots,m}(x_1, \dots , x_m) = \prod_{i=1}^{m} f_i(x_i).$$

- Non-random vectors are constant or deterministic.
- They are also considered random vectors, however, with probability 1 of being equal to a constant (there’s nothing random going on here).
- They are trivially independent of all other random vectors.



Independent

Correlation coefficients

Expected Value

- **Expected value** of a random vector (or random matrix) is defined to be the vector of expected values of its components.

- Properties of expectation carry over from the univariate case.
- **Definition**

Let X be a p-variate random vector. Then the expected value of X is:

$$\mu_X = E(X) = E[(X_1 \ \cdots \ X_p)^T],$$

and if X is continuous then:

$$E(X) = \int x f(x)\, dx$$

Mean

- The **mean** E(X) is the best constant predictor of X in terms of the mean squared error (MSE):

$$E(X) = \arg\min_{c\in\mathbb{R}^p} E\|X - c\|^2.$$

Proof:

$$\frac{\partial}{\partial x}\left[(X-c)^T(X-c)\right] = \frac{\partial}{\partial x}\left[X^TX - 2c^TX - c^Tc\right] = 2(X-c).$$

Then at minimum differential is zero:

$$E[X-c] = 0 \implies E[X] = c$$

Since the second derivative is positive (equal 2), the solution is unique.

Mean

- Let A be a m × p matrix and Y be an m-variate random vector.

  Then $E(AX + Y) = AE(X) + E(Y)$.

- Let b be a constant vector. Then:

  $E(b^T X) = b^T E(X)$.

Covariance Matrix

- Let X be a p-variate random vector. The covariance matrix of X is defined as:

$$\Sigma_{XX} = Var(X) = E\{(X-\mu)^T(X-\mu)\}$$

$$= \begin{pmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_p) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p,X_1) & Cov(X_p,X_2) & \dots & Var(X_p) \end{pmatrix}.$$

Where:

$$Var(X_i) = E\{(X_i - \mu(X_i)) \cdot (X_i - \mu(X_i))\}$$

$$Cov(X_i, Y_j) = E\{(X_i - \mu(X_i)) \cdot (Y_j - \mu(Y_j))\}$$

Cross-covariance matrix

We define the covariance matrix (cross-covariance) between X and Y to be

$$\Sigma_{XY} = Cov(X, Y) = E\{(X - \mu_X)^T (Y - \mu_Y)\}$$

$$= \begin{pmatrix} Cov(X_1, Y_1) & Cov(X_1, Y_2) & \dots & Cov(X_1, Y_m) \\ Cov(X_2, Y_1) & Cov(X_2, Y_2) & \dots & Cov(X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p, Y_1) & Cov(X_p, Y_2) & \dots & Cov(X_p, Y_m) \end{pmatrix} .$$

Some relations:

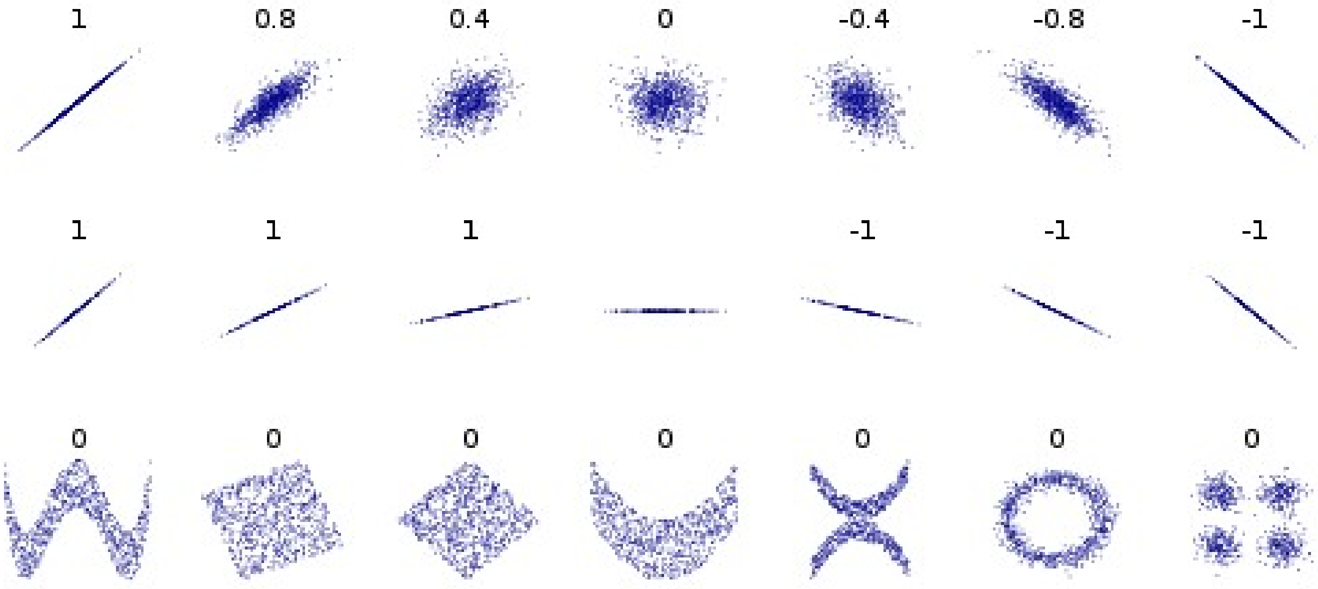
$$Cov(AX, BY) = ACov(X, Y)B^T$$

$$Cov(X + a, Y + b) = Cov(X, Y)$$

Covariance Matrix

Let A be an m × p matrix. Let Y be a random vector. Then

- Var(AX) = A Var(X)A^T .
- Var(X + Y) = Var(X) + Var(Y) + 2 Cov(X, Y).
- If X and Y are independent then Cov(X, Y) = 0.
- Opposite is not true, if Cov(X, Y) = 0 then X and Y are note necessary independent.



Cov(X,Y)=0
and dependent.

Standard deviation

- Standard deviation σ (sigma) is the square root of the variance of X; it is the square root of the average value of $(X - \mu)^2$.
- Let X be a random variable with mean value μ:

$$\mathrm{E}[X] = \mu.$$

E denotes the average or expected value of X.

$$\begin{aligned}
\sigma &= \sqrt{\mathrm{E}[(X-\mu)^2]} \\
&= \sqrt{\mathrm{E}[X^2] + \mathrm{E}[-2\mu X] + \mathrm{E}[\mu^2]} \\
&= \sqrt{\mathrm{E}[X^2] - 2\mu\,\mathrm{E}[X] + \mu^2} \\
&= \sqrt{\mathrm{E}[X^2] - 2\mu^2 + \mu^2} \\
&= \sqrt{\mathrm{E}[X^2] - \mu^2} \\
&= \sqrt{\mathrm{E}[X^2] - (\mathrm{E}[X])^2}
\end{aligned}$$

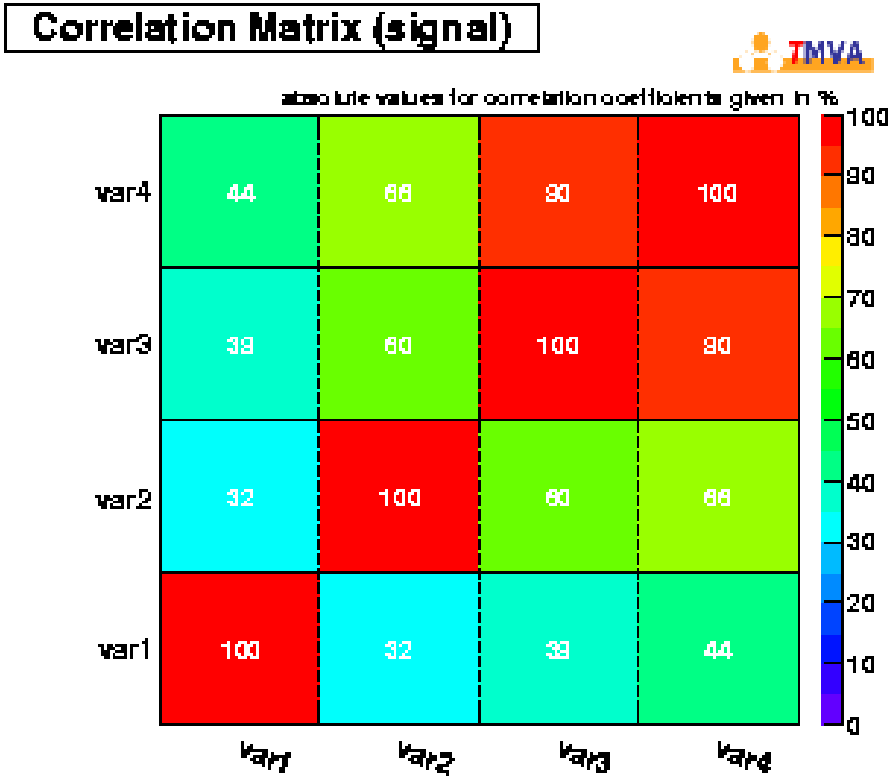
- Standard deviation is a measure of the amount of variation or dispersion of a set of values.
- It is commonly used to measure confidence in statistical conclusions.
- In science we report the standard deviation of experimental data, and only effects more than two standard deviations away from a null expectation are considered statistically significant

Correlation coefficient

- Linear correlation coefficient:

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

where σ(X) is the standard deviation. The correlation coefficient ranges from – 1 (anti-correlation) to + 1 (fully correlated).



	var1	var2	var3	var4
var4	44	68	90	100
var3	39	60	100	90
var2	32	100	60	68
var1	100	32	39	44

Linear correlation between variables (program TMVA used in High Energy Physics)

- Variables are 100% correlated with themselves.
- Var3 and Var4 are strongly correlated
- Array is symmetric
-

Covariance matrix

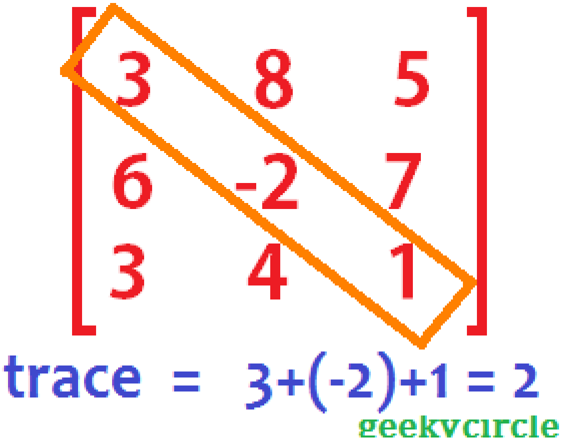
- Let X and Y be p- and m-variate vectors. Then

$$Var\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{pmatrix}.$$

- Remark

  The off-diagonal blocks of the covariance matrix are cross-covariances.

Trace and variance



$$\begin{bmatrix} 3 & 8 & 5 \\ 6 & -2 & 7 \\ 3 & 4 & 1 \end{bmatrix}$$

trace = 3+(-2)+1 = 2

Let A = (a_{ij}) be a square matrices of dimension d × d.

- The trace of A is the sum of its diagonal elements:

$$tr(A) = \sum_i a_{ii}$$

- The mean is the best constant predictor of X in terms of the MSE (shown already):

$$E(X) = \arg\min_{c \in \mathbb{R}^p} E\|X - c\|^2$$

- The **total variance of X** is defined as the MSE of the mean:

$$\mathrm{E} \mathrel{||} \mathrm{X} - \mathrm{E(X)} \mathrel{||}^2 = \sum_{i=1}^{N} E(X_i - E(X_i))^2 = {}_{i=1}^{N} Var(X_i) = tr(Var(X))$$

- The **total variance of X** measures the overall variability of the components of X around the mean E(X). The standard deviation is the commonly used measure of variability.