High Dimensional Data Analysis

Chapter 1: Multivariate Normal Distribution

Outline

- Random Vector
- Multivariate Normal Distribution
- Properties of Multivariate Normal Distribution
- Estimation of the mean vector and Covariance Matrix
- Multiple and Partial Correlation Coefficients

Backgroud

- The generalization of the univariate normal distribution to multiple variables is called the multivariate normal distribution (MVN).
- Many multivariate techniques rely on this distribution in some manner.
- Although real data may not come from a true MVN, the MVN provides a robust approximation, and has many nice mathematical properties.
- Furthermore, because of the central limit theorem, many multivariate statistics converge to the MVN distribution as the sample size increases.

Random Events and Random Variables

- **Random Event**: An outcome of an experiment that cannot be predicted with certainty.
  - Example: It might rain this afternoon, or you might roll a 4 on a die.
- **Random Variable**: A variable whose value depends on the outcome of a probabilistic experiment.
- **Examples**:
  - Roll a die: Let X be the number shown (X could be 1, 2, 3, 4, 5, or 6).
  - Check the weather: Let Y be the temperature in °C (e.g., Y could be 20.5 or 25.3).
- **Notation**: We use capital letters like X or Y for random variables.

Types and Properties of Random Variables

- **Two Types of Random Variables**:
  - *Discrete*: Takes specific, countable values.
    - Example: Number of students in a class ($X = 20, 21, 22, \dots$).
  - *Continuous*: Takes any value in a range.
    - Example: Time to finish a race ($Y = 10.2$ seconds, 10.25 seconds, etc.).
- **Key Ideas**:
  - *Probability*: Tells us how likely each value is.
    - Example: For a die, $P(X = 6) = \frac{1}{6}$.
  - *Average (Expected Value)*: The "typical" value.
    - Example: For a die, the average roll is about 3.5.
  - *Spread (Variance)*: How much the values vary.

What is a Random Vector?

- **Definition**: A random vector is a group of random variables combined into one object.
- **Notation**: Written as $\mathbf{X} = (x_1, x_2, \ldots, x_p)$, where each x_i is a random variable.
- **Examples**:
  - Measure a student's height (x_1) and weight (x_2): $\mathbf{X} = (x_1, x_2)$.
  - Record scores in math (x_1), science (x_2), and history (x_3): $\mathbf{X} = (x_1, x_2, x_3)$.

Covariance Matrix

- **Covariance Matrix**:

$$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] =$$

$$\begin{pmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \cdots & \mathrm{Cov}(x_1, x_p) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Var}(x_2) & \cdots & \mathrm{Cov}(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_p, x_1) & \mathrm{Cov}(x_p, x_2) & \cdots & \mathrm{Var}(x_p) \end{pmatrix}$$

- For any x_i and x_j, the covariance is:

$$\mathrm{Cov}(x_i, x_j) = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])],$$

obviously, $\mathbf{\Sigma}$ is symmetric.

Properties of the Mean Vector

- **Linearity Property**: For a constant matrix A and vector $\mathbf{X}$

$$E(A\mathbf{X}) = AE(\mathbf{X}) = A\boldsymbol{\mu},$$

where $E(\mathbf{X}) = \boldsymbol{\mu}$.

- **Example**: If $\boldsymbol{\mu} = \begin{pmatrix} 170 \\ 65 \end{pmatrix}$ and $A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$, then:

$$E(A\mathbf{X}) = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 170 \\ 65 \end{pmatrix} = \begin{pmatrix} 405 \\ 365 \end{pmatrix}$$

Properties of the Covariance Matrix

- **Variance Property**: For a random vector $\mathbf{X}$ and matrix A, the variance of $A\mathbf{X}$ is:

$$\text{Var}(A\mathbf{X}) = AD(\mathbf{X})A^\top,$$

  where $D(\mathbf{X})$ is the covariance matrix of $\mathbf{X}$.
- **Interpretation**: Scaling $\mathbf{X}$ with A affects the spread of the data.
- **Covariance Property**: For random vectors $\mathbf{X}$ and $\mathbf{Y}$, the constant matrices A and B

$$\text{Cov}(A\mathbf{X}, B\mathbf{Y}) = A\text{Cov}(\mathbf{X}, \mathbf{Y})B^\top,$$

Non-Negative Definiteness

- A covariance matrix is **non-negative definite**, meaning for any vector $v \neq 0$, $v^T \boldsymbol{\Sigma} v \geq 0$.
- Proof: For any vector $v \neq 0$,

$$\begin{aligned} v^T \boldsymbol{\Sigma} v &= v^T \mathbb{E}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T v \\ &= \mathbb{E}[v^T (\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T v] \\ &= \mathbb{E}[||(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T v||^2] \geq 0, \end{aligned}$$

  since the expectation of a squared term is always non-negative.
- This holds for all covariance matrices, regardless of the random variables' dependencies.

Positive Definiteness

- **Assumption**: We often assume $\boldsymbol{\Sigma}$ is **positive definite** ($v^T \boldsymbol{\Sigma} v > 0$ for all $v \neq 0$) to exclude degenerate cases (e.g., linearly dependent variables).
- A covariance matrix is positive definite only if $\boldsymbol{X}$ has no linearly dependent components (full rank).
- **Why assume positive definiteness?**
  - Ensures $\boldsymbol{\Sigma}$ is invertible, critical for multivariate normal distributions and optimization (e.g., Gaussian processes).
  - Improves numerical stability in computations.
  - Small perturbations (e.g., adding $\epsilon \boldsymbol{I}$) can enforce positive definiteness.
- Positive definite $\boldsymbol{\Sigma}$ has all positive eigenvalues, ensuring non-degeneracy.

Univariate Normal Distribution

A random variable x is normally distributed with mean μ and standard deviation $\sigma > 0$ can be written in the following notation:

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

The probability density function of univariate normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- The mean is μ
- The variance is σ^2
- The standard deviation is σ
- Standard notation for normal distributions will be extended for the MVN distribution.

Multivariate Normal Distribution

The multivariate normal distribution of a p-dimensional random vector:

$$\boldsymbol{X} = [x_1, \ldots, x_p]^\top$$

can be written in the following notation:

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with p-dimensional mean vector:

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] = \begin{bmatrix} \mathbb{E}[x_1] \\ \vdots \\ \mathbb{E}[x_p] \end{bmatrix} \in \mathbb{R}^p$$

and covariance matrix:

$$\boldsymbol{\Sigma} = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right] \in \mathbb{R}^{p \times p}.$$

Multivariate Normal Distribution

The density function of univariate normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where μ is the mean and σ^2 is the variance with $\sigma > 0$.
The density function of non-singular p-dimensional multivariate normal distribution is:

$$f(x) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean and $\boldsymbol{\Sigma}$ is the $p \times p$ (non-singular) covariance matrix.

Density Function

Suppose the p-dimensional random vector $\boldsymbol{x}$ has the density function:

$$f(\boldsymbol{x}) = K \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{b})^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{b})\right),$$

where $K \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^p$ and $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ is symmetric positive definite. Then:

$$K = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}}, \quad \boldsymbol{b} = \boldsymbol{\mu} \quad \text{and} \quad \boldsymbol{A} = \boldsymbol{\Sigma}^{-1}.$$

Multivariate Normal Distribution

If the density of a p-dimensional random vector $\boldsymbol{x}$ is:

$$K \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{b})^{\top}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{b})\right),$$

where $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ is symmetric positive definite, then:

$$\mathbb{E}[x] = \boldsymbol{b} \text{ and } Cov[x] = \boldsymbol{A}^{-1}.$$

Conversely, given a vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and a positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, there is a multivariate normal density:

$$f(\boldsymbol{X} \mid \mu, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\mu)^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\mu)\right).$$

Visualizing the MVN is difficult for more than two dimensions, so I will demonstrate some plots with two variables - the bivariate normal distribution.

Bivariate Normal Distribution

We consider the (non-singular) bivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where:

$$\boldsymbol{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

We have:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}.$$

Bivariate Normal Density

Let $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, then we have $\det(\boldsymbol{\Sigma}) = \sigma_{11}\sigma_{22}(1-\rho^2)$ and:

$$(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \frac{1}{1-\rho^2}\Big[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\Big]. \tag{1}$$

The density function is:

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} \exp\Big[-\frac{1}{2(1-\rho^2)}\Big(\frac{(x_1-\mu_1)^2}{\sigma_{11}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}}\Big)\Big].$$

If $\rho = 0$, then the variables x_1 and x_2 are independent.

Probability Density Contour

The density of a p-dimensional normal variable:

$$n(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right)$$

implies the multivariate normal density is constant on surfaces where the square of the distance $(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$ is a constant.
These paths are called contours:

$$\{\boldsymbol{X} : (\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = c^2\} = \text{surface of an hyperellipsoid centered at } \boldsymbol{\mu},$$

where $c > 0$ is a fixed constant.

Probability Density Contour

Consider the hyperellipsoid with surface defined by $\boldsymbol{X} \in \mathbb{R}^p$ such that:

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = c^2.$$

Denote the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$ by:

$$(\lambda_1, u_1), (\lambda_2, u_2), \ldots, (\lambda_p, u_p),$$

then the hyperellipsoid is centered at μ and have axes (vertices):

$$\pm c\sqrt{\lambda_1} u_1, \quad \pm c\sqrt{\lambda_2} u_2, \ldots, \quad \pm c\sqrt{\lambda_p} u_p.$$

Probability Density Contour

For bivariate normal distribution with $\sigma_{11} = \sigma_{22}$, we have:

$$\lambda_1 = \sigma_{11} + \sigma_{12}, \quad u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \lambda_2 = \sigma_{11} - \sigma_{12} \quad \text{and} \quad u_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

If we additionally suppose $\sigma_{12} > 0$, it leads to the following figure:

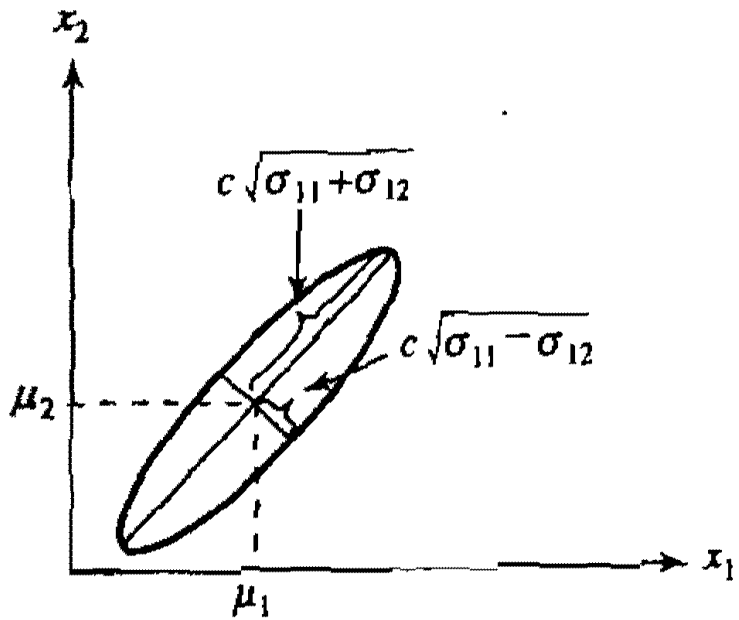


Figure: Probability density contour of bivariate normal distribution

Properties of Multivariate Normal Distribution

Some properties of normally distributed variables:

1. The linear transform of multivariate normal variates are normally distributed.
2. The marginal distributions derived from multivariate normal distributions are also normal distributions.
3. The conditional distributions derived from multivariate normal distributions are also normal distributions.

Properties of Multivariate Normal Distribution

Property

Let $X \sim N_p(\mu, \Sigma)$, B be an $s \times p$ constant matrix, and d be an s-dimensional constant vector. Let $Y = BX + d$, then $Y \sim N_s(B\mu + d, B\Sigma B^T)$.

This property shows that any linear combination of a normal random vector still follows a normal distribution.

Properties of Multivariate Normal Distribution

Corollary

Let $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}^{r}_{p-r} \sim N_p(\mu, \Sigma)$, *and partition* μ, Σ *as*

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}^{r}_{p-r}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{r}_{p-r},$$

then $X^{(1)} \sim N_r(\mu^{(1)}, \Sigma_{11})$, *and* $X^{(2)} \sim N_{p-r}(\mu^{(2)}, \Sigma_{22})$.

This corollary shows that the marginal distributions of a multivariate normal distribution are still normal distributions. However, the converse is not necessarily true.

Properties of Multivariate Normal Distribution

Example 1

Let $X \sim N_3(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 3 \\ 0 \\ -2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 3 \\ 0 & 3 & 4 \end{pmatrix},$$

find the distribution of $Y = BX + d$, where

$$B = \begin{pmatrix} 2 & 0 & -1 \\ 9 & 3 & -8 \end{pmatrix}, \quad d = \begin{pmatrix} -3 \\ 7 \end{pmatrix}.$$

Solution

- *Mean of Y:* $\mathbb{E}[Y] = B\mu + d$.

$$\mathbb{E}[Y] = \begin{pmatrix} 2 & 0 & -1 \\ 9 & 3 & -8 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \\ -2 \end{pmatrix} + \begin{pmatrix} -3 \\ 7 \end{pmatrix} = \begin{pmatrix} 8 \\ 43 \end{pmatrix} + \begin{pmatrix} -3 \\ 7 \end{pmatrix} = \begin{pmatrix} 5 \\ 50 \end{pmatrix}.$$

- *Covariance of Y:* $Cov(Y) = B\Sigma B^T$.

$$Cov(Y) = \begin{pmatrix} 2 & 0 & -1 \\ 9 & 3 & -8 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 3 \\ 0 & 3 & 4 \end{pmatrix} \begin{pmatrix} 2 & 9 \\ 0 & 3 \\ -1 & -8 \end{pmatrix} = \begin{pmatrix} 6 & -11 \\ -37 & 198 \end{pmatrix}$$

- *Thus,* $Y \sim \mathcal{N}_2\left(\begin{pmatrix} 5 \\ 50 \end{pmatrix}, \begin{pmatrix} 6 & -11 \\ -37 & 198 \end{pmatrix}\right)$.

Properties of Multivariate Normal Distribution

Theorem

Let

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}_{p-r}^{r} \sim N_p \left(\begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

then $X^{(1)}$ *and* $X^{(2)}$ *are independent if and only if* $\Sigma_{12} = 0$.

Property

Let $X_1, \cdots, X_k$ *be mutually independent, and* $X_i \sim N_p(\mu_i, \Sigma_i), i = 1, \cdots, k$, *and* $c_1, \cdots, c_k$ *be arbitrary constants. Then*

$$\sum_{i=1}^{k} c_i X_i \sim N_p \left(\sum_{i=1}^{k} c_i \mu_i, \sum_{i=1}^{k} c_i^2 \Sigma_i \right).$$

Property

Let $X \sim N(\mu, \Sigma)$, *and* $\Sigma > 0$, *then* $(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2(p)$.

Properties of Multivariate Normal Distribution

Theorem

Let $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}^{r}_{p-r} \sim N_p(\mu, \Sigma)$, *with* $\Sigma > 0$, *and partition* μ, Σ *as*

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}^{r}_{p-r}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{r}_{p-r}.$$

Then, given $X^{(2)} = x^{(2)}$, *the conditional distribution of* $X^{(1)}$ *is*

$$X^{(1)}|X^{(2)} = x^{(2)} \sim N_r(\mu_{1,2}, \Sigma_{11,2}),$$

where

$$\mu_{1,2} = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)}), \quad \Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Properties of Multivariate Normal Distribution

Example 2

Let $X = (x_1, x_2, x_3) \sim N_3(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 16 & -4 & 2 \\ -4 & 4 & -1 \\ 2 & -1 & 4 \end{pmatrix}.$$

Find the conditional distribution of

$$\begin{pmatrix} x_2 - x_3 \\ x_1 \end{pmatrix}$$

given $x_1 + 2x_3$.

Solution

Let $y_1 = \begin{pmatrix} x_2 - x_3 \\ x_1 \end{pmatrix}, y_2 = x_1 + 2x_3,$ *then*

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_2 - x_3 \\ x_1 \\ x_1 + 2x_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$E \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix}$$

$$V \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 16 & -4 & 2 \\ -4 & 4 & -1 \\ 2 & -1 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} 10 & -6 & -16 \\ -6 & 16 & 20 \\ -16 & 20 & 40 \end{pmatrix}$$

Solution

The conditional mean and conditional covariance matrix given y_2 are, respectively,

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} -16 \\ 20 \end{pmatrix} \frac{1}{40}(y_2 + 3) = \begin{pmatrix} -\frac{2}{5}y_2 + \frac{4}{5} \\ \frac{1}{2}y_2 + 2\frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{2}{5}x_1 - \frac{4}{5}x_3 + \frac{4}{5} \\ \frac{1}{2}x_1 + x_3 + 2\frac{1}{2} \end{pmatrix}$$

$$\begin{pmatrix} 10 & -6 \\ -6 & 16 \end{pmatrix} - \begin{pmatrix} -16 \\ 20 \end{pmatrix} \frac{1}{40}(-16, 20) = \begin{pmatrix} 3\frac{3}{5} & 2 \\ 2 & 6 \end{pmatrix}$$

Therefore,

$$\begin{pmatrix} x_2 - x_3 \\ x_1 \end{pmatrix} | x_1 + 2x_3 \sim N_2 \left[\begin{pmatrix} -\frac{2}{5}x_1 - \frac{4}{5}x_3 + \frac{4}{5} \\ \frac{1}{2}x_1 + x_3 + \frac{5}{2} \end{pmatrix}, \begin{pmatrix} \frac{18}{5} & 2 \\ 2 & 6 \end{pmatrix} \right]$$

Estimation of Mean Vector and Covariance Matrix

Sample Mean Vector and Sample Covariance Matrix

Let $X_{(i)} = (x_{i1}, \cdots, x_{ip})$ $(i = 1, \cdots, n)$ be a random sample from the population $X = (x_1, \cdots, x_p) \sim N_p(\mu, \Sigma)$, with $\Sigma > 0$. Denote the sample data matrix as

$$M = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} X'_{(1)} \\ \vdots \\ X'_{(n)} \end{pmatrix}.$$

Next, we introduce the sample mean vector, sample deviation matrix, sample covariance matrix, and sample correlation matrix.

Estimation of Mean Vector and Covariance Matrix

The sample mean vector is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n} M^T I_n,$$

where

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad i = 1, \cdots, p,$$

is the sample mean of x_i.

Estimation of Mean Vector and Covariance Matrix

The sample deviation matrix is

$$A = \sum_{i=1}^{n}(X_{(i)} - \overline{X})(X_{(i)} - \overline{X})^T = M^T M - n\overline{X} \cdot \overline{X}^T$$

$$= M^T \left[I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right] M = (a_{ij})_{p\times p},$$

where

$$a_{ij} = \sum_{l=1}^{n}(x_{il} - \overline{x}_i)(x_{lj} - \overline{x}_j), \quad i, j = 1, \cdots, p.$$

Estimation of Mean Vector and Covariance Matrix

The sample covariance matrix is

$$S = \frac{1}{n-1}A = (s_{ij})_{p\times p},$$

where

$$s_{ij} = \frac{1}{n-1}\sum_{l=1}^{n}(x_{il} - \bar{x}_i)(x_{lj} - \bar{x}_j), \quad i,j = 1,\cdots,p.$$

s_{ii} is called the sample variance of variable x_i, and its arithmetic square root $\sqrt{s_{ii}}$ is called the sample standard deviation of variable x_i.
The sample correlation matrix is

$$R = (r_{ij})_{p\times p},$$

where

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} = \frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}, \quad i,j = 1,\cdots,p.$$

Estimation of Mean Vector and Covariance Matrix

Theorem

Let $\overline{X}$ and A be the sample mean vector and sample deviation matrix from a population $N_p(\mu, \Sigma)$, respectively. Then:

1. $$\overline{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right);$$
2. *A can be expressed as $A = \sum_{i=1}^{n-1} Z_i Z_i^T$, where $Z_1, \cdots, Z_{n-1}$ are independent and identically distributed as $N_p(0, \Sigma)$;*
3. *$\overline{X}$ and A are mutually independent.*

Estimation of Mean Vector and Covariance Matrix

Maximum Likelihood Estimation and Its Properties

Let $X_{(i)} = (x_{i1}, \cdots, x_{ip})^T$ $(i = 1, \cdots, n)$ be a random sample from the normal distribution $N_p(\mu, \Sigma)$, with $\Sigma > 0$. In this section, we discuss the maximum likelihood estimation of the parameters μ and Σ.

For this purpose, we consider the joint density function of the sample

$$L(\mu, \Sigma) = \prod_{i=1}^{n} f(x_{(i)}, \mu, \Sigma)$$

as a function of μ and Σ, and call it the likelihood function, i.e.,

Estimation of Mean Vector and Covariance Matrix

$$L(\mu,\Sigma)=\prod_{i=1}^{n}\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x_{(i)}-\mu)^T\Sigma^{-1}(x_{(i)}-\mu)\right\}$$

$$=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_{(i)}-\mu)^T\Sigma^{-1}(x_{(i)}-\mu)\right\}$$

$$=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}[(x_{(i)}-\mu)^T\Sigma^{-1}(x_{(i)}-\mu)]\right\}$$

$$=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}[\Sigma^{-1}(x_{(i)}-\mu)(x_{(i)}-\mu)^T]\right\}$$

$$=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left\{\mathrm{tr}\left(-\frac{1}{2}\Sigma^{-1}\sum_{i=1}^{n}(x_{(i)}-\mu)(x_{(i)}-\mu)^T\right)\right\}.$$

Estimation of Mean Vector and Covariance Matrix

Since

$$\sum_{i=1}^{n}(x_{(i)}-\mu)(x_{(i)}-\mu)^T$$

$$=\sum_{i=1}^{n}(x_{(i)}-\overline{X}+\overline{X}-\mu)(x_{(i)}-\overline{X}+\overline{X}-\mu)^T$$

$$=\sum_{i=1}^{n}(x_{(i)}-\overline{X})(x_{(i)}-\overline{X})^T+n(\overline{X}-\mu)(\overline{X}-\mu)^T$$

$$=A+n(\overline{X}-\mu)(\overline{X}-\mu)^T,$$

we have

$$L(\mu,\Sigma)=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}[A+n(\overline{X}-\mu)(\overline{X}-\mu)^T]\right]\right\},$$

Estimation of Mean Vector and Covariance Matrix

$$\frac{\partial \ln L(\mu, \Sigma)}{\partial \mu} = n\Sigma^{-1}(\overline{X} - \mu),$$

$$\frac{\partial \ln L(\overline{X}, \Sigma)}{\partial \Sigma^{-1}} = n\Sigma - \frac{n}{2}\text{diag}(\Sigma) - A + \frac{1}{2}\text{diag}(A).$$

Set

$$\frac{\partial \ln L(\mu, \Sigma)}{\partial \mu} = 0, \quad \frac{\partial \ln L(\overline{X}, \Sigma)}{\partial \Sigma^{-1}} = 0,$$

we obtain the maximum likelihood estimates of μ and Σ as

$$\hat{\mu} = \bar{X}, \quad \hat{\Sigma} = \frac{1}{n}A,$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}, \quad A = \sum_{i=1}^{n}(X_{(i)} - \bar{X})(X_{(i)} - \bar{X})^T.$$

Estimation of Mean Vector and Covariance Matrix

Since

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu,$$

$\bar{X}$ is an unbiased estimator of μ.
We have

$$E(A) = E\left(\sum_{i=1}^{n-1} Z_i Z_i^T\right) = \sum_{i=1}^{n-1} E(Z_i Z_i^T) = \sum_{i=1}^{n-1} \mathrm{Var}(Z_i) = \sum_{i=1}^{n-1} \Sigma = (n-1)\Sigma,$$

therefore, the maximum likelihood estimator of Σ is not unbiased.
However,

$$S = \frac{1}{n-1}A = \sum_{i=1}^{n}(X_{(i)} - \overline{X})(X_{(i)} - \overline{X})^T = \frac{n}{n-1}\hat{\Sigma}$$

is an unbiased estimator of Σ. Usually, $\overline{X}$ is called the sample mean, and S is called the sample covariance matrix.

Multiple and Partial Correlation Coefficients

Multiple Correlation Coefficient

- The (simple) correlation coefficient measures the strength of the linear relationship between one random variable x and another random variable y.
- The multiple correlation coefficient measures the strength of the linear relationship between one random variable y and a set of random variables $x_1, x_2, \ldots, x_p$.

Let

$$E\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \quad V\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \sigma_{yy} & \sigma'_{xy} \\ \sigma_{xy} & \Sigma_{xx} \end{pmatrix},$$

The correlation matrix of $\begin{pmatrix} y \\ x \end{pmatrix}$ is

$$= \begin{pmatrix} 1 & \rho'_{xy} \\ \rho_{xy} & R_{xx} \end{pmatrix}$$

Multiple and Partial Correlation Coefficients

Then the maximum correlation coefficient between y and a linear function of x, $l'x(l \neq 0)$, is called the multiple correlation coefficient between y and x, denoted by $\rho_{y\cdot x}$ or $\rho_{y\cdot 1,2,\cdots,p}$. It measures the degree of correlation between one variable y and a set of variables $x_1, x_2, \cdots, x_p$.

$$\rho_{y\cdot x} = \max_{l\neq 0} \rho(y, l'x) = \rho(y, \sigma'_{xy}\Sigma_{xx}^{-1}x)$$

$$= \sqrt{\frac{\sigma'_{xy}\Sigma_{xx}^{-1}\sigma_{xy}}{\sigma_{yy}}} = \sqrt{\rho'_{xy}R_{xx}^{-1}\rho_{xy}}$$

If $x_1, x_2, \cdots, x_p$ are mutually uncorrelated, then

$$\rho^2_{y\cdot x} = \rho'_{xy}R_{xx}^{-1}\rho_{xy} = \rho'_{xy}\rho_{xy}$$

$$= \rho^2(y, x_1) + \cdots + \rho^2(y, x_p)$$

Multiple and Partial Correlation Coefficients

Maximum Likelihood Estimation of $\rho_{y \cdot x}$

Let the sample covariance matrix be $V \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} s_{yy} & s'_{xy} \\ s_{xy} & S_{xx} \end{pmatrix}$, and the sample correlation matrix of $\begin{pmatrix} y \\ x \end{pmatrix}$ be $= \begin{pmatrix} 1 & r'_{xy} \\ r_{xy} & R_{xx} \end{pmatrix}$.

Here $n > p$, then under the assumption of multivariate normality, the maximum likelihood estimate of the multiple correlation coefficient $\rho_{y \cdot x}$ is

$$r_{y \cdot x} = \sqrt{\frac{s'_{xy} S_{xx}^{-1} s_{xy}}{s_{yy}}} = \sqrt{r'_{xy} R_{xx}^{-1} r_{xy}}$$

which is called the sample multiple correlation coefficient.

Multiple and Partial Correlation Coefficients

Partial Correlation Coefficient

- The correlation between two variables is often influenced not only by each other but also by a series of other variables.
- For this reason, the correlation coefficient is sometimes called the total (or gross) correlation coefficient, meaning it includes the correlation brought by all influences.
- The correlation coefficient is also sometimes called the simple correlation coefficient or Pearson correlation coefficient or zero-order partial correlation coefficient.

Multiple and Partial Correlation Coefficients

Example 3

A physical test was conducted on 31 individuals. The seven indicators observed or tested are: Age (x_1), Weight (x_2), Vital Capacity (x_3), Time for 1.5 Mile Run (x_4), Pulse at Rest (x_5), Pulse while Running (x_6), and Maximum Pulse Recorded while Running (x_7). The data are listed in the following table.

- The sample multiple correlation coefficient between x_3 and $x_1, x_2, x_4, x_5, x_6, x_7$ can be calculated as:

$$R_{3\cdot 1,2,4,5,6,7} = 0.9209$$

ID	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	44	89.47	44.609	11.37	62	178	182
2	40	75.07	45.313	10.07	62	185	185
3	44	85.84	54.297	8.65	45	156	168
4	42	68.15	59.571	8.17	40	166	172
5	38	89.02	49.874	9.22	55	178	180
6	47	77.45	44.811	11.63	58	176	176
7	40	75.98	45.681	11.95	70	176	180
8	43	81.19	49.091	10.85	64	162	170
9	44	81.42	39.442	13.08	63	174	176
10	38	81.87	60.055	8.63	48	170	186
11	44	73.03	50.541	10.13	45	168	168
12	45	87.66	37.388	14.03	56	186	192
13	45	66.45	44.754	11.12	51	176	176
14	47	79.15	47.273	10.6	47	162	164
15	54	83.12	51.855	10.33	50	166	170
16	49	81.42	49.156	8.95	44	180	185
17	51	69.63	40.836	10.95	57	168	172
18	51	77.91	46.672	10.0	48	162	168
19	48	91.63	46.774	10.25	48	162	164
20	49	73.37	50.388	10.08	76	168	168
21	57	73.37	39.407	12.63	58	174	176
22	54	79.38	46.08	11.17	62	156	165
23	52	76.32	45.441	9.63	48	164	166
24	50	70.87	54.625	8.92	48	146	155
25	51	67.25	45.118	11.08	48	172	172
26	54	91.63	39.203	12.88	44	168	172
27	51	73.71	45.79	10.47	59	186	188
28	57	59.08	50.545	9.93	49	148	155
29	49	76.32	48.673	9.4	56	186	188
30	48	61.24	47.92	11.5	52	170	176
31	52	82.78	47.467	10.5	53	170	172

Python code for $R_{3;1,2,4,5,6,7}$

```python
import pandas as pd
import numpy as np
import statsmodels.api as sm

# Partial data (first few rows)
data = {
    'ID': [1, 2, 3, ...],
    'x1': [44, 40, 44, ...],
    'x2': [89.47, 75.07, 85.84,
        ...],
    'x3': [44.609, 45.313,
        54.297, ...],
    'x4': [11.37, 10.07, 8.65,
        ...],
    'x5': [62, 62, 45, ...],
    'x6': [178, 185, 156, ...],
    'x7': [182, 185, 168, ...]
}

# Create DataFrame
df = pd.DataFrame(data)
```

```python
# Predictors
X = df[['x1', 'x2', 'x4', 'x5',
    'x6', 'x7']]

# Dependent variable
y = df['x3']

# Add constant
X = sm.add_constant(X)

# Fit OLS model
model = sm.OLS(y, X).fit()

# R-squared
r_squared = model.rsquared

# Multiple correlation
R = np.sqrt(r_squared)

# Print result
print(f"R_{{3;1,2,4,5,6,7}} =
    {R:.4f}")
```

Multiple and Partial Correlation Coefficients

Partial Correlation Coefficient

- The correlation between two variables is often influenced not only by each other but also by a series of other variables.
- For this reason, the correlation coefficient is sometimes called the total (or gross) correlation coefficient, meaning it includes the correlation brought by all influences.
- The correlation coefficient is also sometimes called the simple correlation coefficient or Pearson correlation coefficient or zero-order partial correlation coefficient.

Multiple and Partial Correlation Coefficients

Example 4

x_1–household food expenditure

x_2–household clothing expenditure

x_3–household income

- There exists a strong positive correlation between x_1 and x_2.

- The strong positive correlations of x_3 with x_1 and x_2 respectively lead to the strong positive correlation between x_1 and x_2.

- If we can eliminate the influence of x_3 in some way, or control for x_3 (i.e., keep x_3 constant), then the correlation between x_1 and x_2 (reflecting the net relationship) might be quite different, and could very likely show a negative correlation.

Multiple and Partial Correlation Coefficients

Partition $x, \Sigma(>0), S$ as follows:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{matrix} k \\ p-k \end{matrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{matrix} k \\ p-k \end{matrix}, \quad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{matrix} k \\ p-k \end{matrix}$$

- Call $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ the partial covariance matrix of x_1 given x_2.
- Denote $\Sigma_{11.2} = (\sigma_{ij \cdot k+1,\cdots,p})$, and call $\sigma_{ij \cdot k+1,\cdots,p}$ the partial covariance.
- It is the covariance between x_i and x_j after removing the (linear) influence of $x_2 = (x_{k+1}, \cdots, x_p)'$.

Multiple and Partial Correlation Coefficients

The partial correlation coefficient between x_i and x_j given x_2 is defined as

$$\rho_{ij\cdot k+1,\cdots,p} = \frac{\sigma_{ij\cdot k+1,\cdots,p}}{\sqrt{\sigma_{ii\cdot k+1,\cdots,p}\sigma_{jj\cdot k+1,\cdots,p}}}, \quad 1 \leq i,j \leq k$$

where $\Sigma_{11.2} = (\sigma_{ij\cdot k+1,\cdots,p})$.
- $\rho_{ij\cdot k+1,\cdots,p}$ measures the strength of the relationship between x_i and x_j after removing the (linear) influence of $x_{k+1}, \cdots, x_p$.
- For multivariate normal variables x, since $\Sigma_{11.2}$ is also the conditional covariance matrix, the partial correlation coefficient is the same as the conditional correlation coefficient. Thus, $\rho_{ij\cdot k+1,\cdots,p}$ also measures the strength of the relationship between x_i and x_j given the values of $x_{k+1}, \cdots, x_p$.

Multiple and Partial Correlation Coefficients

- First-order partial correlation coefficients can be directly calculated from correlation coefficients. Let x_1, x_2, x_3 be three random variables, then

$$\rho_{12.3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$$

$$\rho_{13.2} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{23}^2}}$$

$$\rho_{23.1} = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}$$

- (1) $\rho_{12} = 0$ does not imply $\rho_{12.3} = 0$, and vice versa.
- (2) ρ_{12} and $\rho_{12.3}$ are not necessarily of the same sign.
- Furthermore, there is no definitive conclusion on which is larger between ρ_{12} and $\rho_{12.3}$.

Multiple and Partial Correlation Coefficients

General recursive formula for partial correlation coefficients:

$$\rho_{ij\cdot k+1,\cdots,p} = \frac{\rho_{ij\cdot k+2,\cdots,p} - \rho_{i,k+1\cdot k+2,\cdots,p}\rho_{j,k+1\cdot k+2,\cdots,p}}{\sqrt{1-\rho^2_{i,k+1\cdot k+2,\cdots,p}}\sqrt{1-\rho^2_{j,k+1\cdot k+2,\cdots,p}}}$$

Under the assumption of multivariate normality, the maximum likelihood estimate of $\rho_{ij\cdot k+1,\cdots,p}$ is

$$r_{ij\cdot k+1,\cdots,p} = \frac{s_{ij\cdot k+1,\cdots,p}}{\sqrt{s_{ii\cdot k+1,\cdots,p}s_{jj\cdot k+1,\cdots,p}}}, \quad 1 \le i,j \le k$$

where $S_{11.2} = S_{11} - S_{12}S_{22}^{-1}S_{21} = (s_{ij\cdot k+1,\cdots,p})$. $r_{ij\cdot k+1,\cdots,p}$ is called the sample partial correlation coefficient.

Multiple and Partial Correlation Coefficients

Example 5: Suppose measurements of birth weight (ounces), age in days, and diastolic blood pressure (mmHg) were taken for 16 infants. The data are shown in the following table.

No.	Birth Weight x_1	Age (days) x_2	Diastolic BP x_3
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88

Table: Data on birth weight, age, and blood pressure for 16 infants

Multiple and Partial Correlation Coefficients

Solution

$$\hat{R} = \begin{pmatrix} 1 & 0.1068 & 0.4411 \\ 0.1068 & 1 & 0.8708 \\ 0.4411 & 0.8708 & 1 \end{pmatrix}$$

The sample partial correlation coefficient between diastolic blood pressure and birth weight, controlling for age, is

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} = \frac{0.4411 - 0.1068 \times 0.8708}{\sqrt{1 - 0.1068^2}\sqrt{1 - 0.8708^2}} = 0.7121$$

The sample partial correlation coefficient between diastolic blood pressure and age, controlling for birth weight, is

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}} = \frac{0.8708 - 0.1068 \times 0.4411}{\sqrt{1 - 0.1068^2}\sqrt{1 - 0.4411^2}} = 0.9231$$