SF 2930 Regression Analysis

Lecture 2

Multivariate Random Vectors, Multivariate Normal Distribution and Multivariate Normal Random Vectors, Simple Linear Regression in Matrix Terms

Timo Koski

KTH Royal Institute of Technology

19-01- 2023

Learning outcomes

- Random vectors, mean vector, covariance matrix, rules of transformation
- Multivariate normal R.V., rules of transformation
  - Density of a multivariate normal RV
  - Joint PDF of bivariate normal RVs
  - Conditional distributions in a multivariate normal
  - Joint PDF of normal RVs distribution
- Matrix algebra related to Multivariate normal R.V.
  - Standard normal R.V., Rules of transformation,
  - Simple Linear Regression by Random vectors, Likelihood
  - Diagonalization of a Covariance Matrix

PART 0: Euclidean Vector Space $\mathbb{R}^n$, Matrices

Linear Algebra

Remark

These lectures/this course are/is heavily dependent on application of matrix calculus in Euclidean vector spaces. There is a four page matrix calculus Appendix C.1-C.3 in MPV. A more comprehensive presentation and refresher (with several proofs) is Chapter 2 in

Rencher, Alvin C and Schaalje, G Bruce: Linear Models in Statistics*, 2008, John Wiley & Sons*

The results on symmetric, non-negative definite and idempotent matrices and distributions of quadratic forms found in this book are especially useful. Linear Models in Statistics *is digitally available via KTHB.*

Remark

Random Vectors and Multivariate normal distribution are treated, e.g., in Chapters 3 and 4 of
Rencher, Alvin C and Schaalje, G Bruce: Linear Models in Statistics*, 2008, John Wiley & Sons*
and in Chapter 5 of
*Gut, Allan:*An Intermediate Course in Probability. Second Edition*, Springer, 2009*

The statements in this lecture 2 are proved by means of moment generating functions (c.f. the references above) and are omitted, for reasons of time budgeting, here.

Notation : Euclidean vector space $\mathbb{R}^n$

$x_1, x_2, \ldots x_n$ is an n-tuple of real numbers. Then we write

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n \quad \mathbf{x}^T = (x_1, x_2, \ldots x_n) \in \mathbb{R}^n$$

Such $\mathbf{x}$ is said to be a $n \times 1$ vector. $\mathbf{x}^T$ is times a $1 \times n$ vector, the transpose of $\mathbf{x}$.

A Refresher Item: Euclidean vector space $\mathbb{R}^n$, scalar product, norm, distance

A Euclidean vector space is a finite-dimensional inner product (scalar product) space over the real numbers. For $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, the scalar product $\mathbf{x}^T\mathbf{y}$ is defined as

$$\mathbf{x}^T\mathbf{y} := \sum_{i=1}^{n} x_i y_i.$$

$$\mathbf{x}^T\mathbf{y} = (x_1, x_2, \dots x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

A Refresher Item: Euclidean vector space $\mathbb{R}^n$, scalar product, norm, distance

- $$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} y_i x_i = \mathbf{y}^T\mathbf{x}$$
- $$\mathbf{x}^T(\mathbf{y}_1 + \mathbf{y}_2) = \mathbf{x}^T\mathbf{y}_1 + \mathbf{x}^T\mathbf{y}_2$$
- $\mathbf{x}^T\mathbf{x} > 0$ if $\mathbf{x} \neq \mathbf{0}_n$=the $n \times 1$ vector with all n components $= 0$,
  $\mathbf{x}^T\mathbf{x} = 0$ if $\mathbf{x} = \mathbf{0}_n$

  .

EXQ

Show that it follows from the preceding that

$$(\mathbf{y}_1 + \mathbf{y}_2)^T \mathbf{x} = \mathbf{y}_1^T \mathbf{x} + \mathbf{y}_2^T \mathbf{x}$$

Orthogonal and Orthonormal Vectors in the Euclidean space $\mathbb{R}^n$:

Definition

$\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \mathbb{R}^n$ *are called orthogonal, if*

$$\mathbf{x}^T\mathbf{y} = 0.$$

Definition

A set of vectors $\{\mathbf{e}_i\}_{i\in I}$ *is said to be orthonormal, if*

$$\mathbf{e}_i^T\mathbf{e}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

If $\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \mathbb{R}^n$ *are orthogonal, then*

$$\| \mathbf{x} - \mathbf{y} \|^2 = \| \mathbf{x} \|^2 + \| \mathbf{y} \|^2$$

Please check this!

The Standard Basis of Vectors in the Euclidean space $\mathbb{R}^n$:

$$\mathfrak{E}_1 = \begin{pmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \mathfrak{E}_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \mathfrak{E}_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

is an orthonormal set of vectors in $\mathbb{R}^n$ known as the standard basis of $\mathbb{R}^n$. This means that every $\mathbf{x} \in \mathbb{R}^n$ can be uniquely written as

$$\mathbf{x} = \sum_{i=1}^{n} c_i \mathfrak{E}_i,$$

where $c_i = \mathbf{x}^T \mathfrak{E}_i$.

A Refresher Item: Euclidean vector space $\mathbb{R}^n$, norm, distance

We have also a norm $\| \mathbf{x} \|$ in $\mathbb{R}^n$ defined by

$$\| \mathbf{x} \| = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

This norm gives a distance (i.e., a metric) between $\mathbf{x}$ and $\mathbf{y}$ by

$$\| \mathbf{x} - \mathbf{y} \| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$\| \mathbf{x} \|^2 = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}_n \quad \& \quad \| \mathbf{x} - \mathbf{y} \| = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$$

Norms

Remark

To be mathematically precise we are actually here dealing with the so-called l_2- norm $\| \mathbf{x} \|_2$ on $\mathbb{R}^n$. There are other known norms, like $\| \mathbf{x} \|_\infty = \max_i |x_i|$ or the l_p-norm $\| \mathbf{x} \|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$, $p \geq 1$. The norm $\| \mathbf{x} \|_1$ will appear in the regression with Lasso. But since there is at the moment no risk of confusion, we stay with the simpler notation, i.e., $\| \mathbf{x} \|$.

Matrix Addition

A is an $m \times n$ matrix and B is an $m \times n$.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{n2} & \cdots & b_{mn} \end{pmatrix}$$

The matrix sum $A + B$ is

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{12} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{pmatrix}$$

Matrix Multiplication

A is an $m \times n$ matrix and B is an $n \times p$, we say that A and B are conformable for the multiplication AB.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix}$$

The matrix product $C := AB$ is defined to be the $m \times p$ matrix

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{pmatrix}$$

where $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$.

Matrix Multiplication

I.e., the entry c_{ij} of C is found by calculating the scalar product of the ith row vector of A and the jth column vector of B. Therefore, $C = AB$ is

$$C = \begin{pmatrix} a_{11}b_{11} + \cdots + a_{1n}b_{n1} & \cdots & a_{11}b_{1p} + \cdots + a_{1n}b_{np} \\ \vdots & \ddots & \vdots \\ a_{m1}b_{11} + \cdots + a_{mn}b_{n1} & \cdots & a_{m1}b_{1p} + \cdots + a_{mn}b_{np} \end{pmatrix} \quad (1)$$

Matrix Multiplication: A Special Case

$$\mathbf{x}^T\mathbf{y} = (x_1, x_2, \dots x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Multiplication by a Scalar

$$cA = \begin{pmatrix} ca_{11} & ca_{12} & \cdots & ca_{1n} \\ ca_{21} & ca_{22} & \cdots & ca_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{m1} & ca_{m2} & \cdots & ca_{mn} \end{pmatrix}$$

Matrix Multiplication: $\mathbf{x}\mathbf{x}^T$

$\mathbf{x}$ is an $n \times 1$ vector (matrix), $\mathbf{x}^T$ is $1 \times n$. Then $\mathbf{x}\mathbf{x}^T$ is by (1) an $n \times n$ matrix

$$\mathbf{x}\mathbf{x}^T = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} (x_1, x_2, \ldots, x_n) = \begin{pmatrix} x_1^2 & x_1 x_2 & \cdots & x_1 x_n \\ x_2 x_1 & x_2^2 & \cdots & x_2 x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n x_1 & x_n x_2 & \cdots & x_n^2 \end{pmatrix} \tag{2}$$

$$\mathbf{x}^T\mathbf{x} = \sum_{i=1}^{n} x_i^2$$

The trace of a square matrix A, denoted by $\operatorname{Tr} A$, is the sum of its elements of the main diagonal. Hence

$$\operatorname{Tr} \mathbf{x}\mathbf{x}^T = \mathbf{x}^T\mathbf{x} \tag{3}$$

Matrix Multiplication: $\mathbf{y} = A\mathbf{x}$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

$$\mathbf{y} = A\mathbf{x} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ a_{21}x_1 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix}.$$

$\mathbf{y}$ is an $m \times 1$ (matrix) vector.

$$B\mathbf{y} = BA\mathbf{x}, \quad A(\mathbf{x}_1 + \mathbf{x}_2) = A\mathbf{x}_1 + A\mathbf{x}_2$$

The $n \times n$ Identity Matrix

$$\mathbb{I}_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{4}$$

$$\mathbb{I}_n \mathbf{x} = \mathbf{x}, \quad \mathbb{I}_n A = A \quad \text{if } A \text{ is conformable}$$

A System of Linear Equations

The general form of a system of linear equations is

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned} \quad .$$

The system is by the multiplication rule above equivalent with the single matrix equation

$$A\mathbf{x} = \mathbf{b}.$$

If $m = n$ and A has an inverse matrix A^{-1}, i.e., $A^{-1}A = AA^{-1} = \mathbb{I}_n$, then

$$\mathbf{x} = A^{-1}\mathbf{b}.$$

Matrix Transpose: A^T

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad m \times n$$

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix} \quad n \times m$$

$$(AB)^T = B^T A^T, \quad (A+B)^T = A^T + B^T.$$

By conformability: $A^T A$ is $n \times n$ and AA^T is $m \times m$.

Symmetric Matrix

Definition

A matrix A is said to be symmetric, if

$$A^T = A$$

Example

Any diagonal matrix

$$D_n = \begin{pmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & d_n \end{pmatrix} \tag{5}$$

is clearly symmetric. A special case: $\mathbb{I}_n$ in (4).

Inverse of Diagonal Matrix

Example

If all elements d_i on the main diagonal of a diagonal matrix D_n are positive, then the inverse D_n^{-1} exists and is a symmetric matrix given by

$$D_n^{-1} = \begin{pmatrix} 1/d_1 & 0 & 0 & \dots & 0 \\ 0 & 1/d_2 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & 1/d_n \end{pmatrix} \tag{6}$$

Easy to check: $D_n^{-1} D_n = D_n D_n^{-1} = \mathbb{I}_n$

A Matrix of Usefulness/Importance in Regression Analysis

Let us define the $n \times n$ matrix C_{ce} by

$$C_{ce} := \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \quad \text{where} \quad \mathbf{1}_n := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{7}$$

Note that as a special case of (2)

$$\mathbf{1}_n\mathbf{1}_n^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \quad n \times n, \quad \mathbf{1}_n^T\mathbf{1}_n = n$$

The Centering Matrix (is of Usefulness/Importance in Regression Analysis)

Take an $n \times 1$ vector $\mathbf{x}$. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then, since $\mathbf{1}_n^T\mathbf{x} = \sum_{i=1}^{n} x_i$,

$$C_{ce}\mathbf{x} = \mathbb{I}_n\mathbf{x} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{x} = \mathbf{x} - \bar{x}\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} = \begin{pmatrix}x_1 - \bar{x}\\x_2 - \bar{x}\\\vdots\\x_n - \bar{x}\end{pmatrix}. \tag{8}$$

Thus C_{ce} does a centering of $\mathbf{x}$.

EXQ: Symmetry and Idempotency of $\mathbf{C}_{ce}$

Definition

An $n \times n$- matrix A is said to be **idempotent**, *if*

$$A^2 = AA = A$$

Show/check that

- C_{ce} *is symmetric.*
- C_{ce} *is* **idempotent**.

To check that C_{ce} is symmetric should be easy. For idempotency, note that $\mathbf{1}_n^T \mathbf{1}_n = n$.

Orthogonal Projection

Definition

I an $n \times n$- matrix A is

- *symmetric.*
- *and idempotent,*

then it called **an orthogonal projection matrix**

An orthogonal projection matrix splits $\mathbb{R}^n$ into a direct sum of two subspaces, its range space and its null space.

A Quadratic Form

Since $\mathbf{C}_{ce}$ is idempotent and symmetric,

$$\mathbf{x}^T C_{ce}\mathbf{x} = \mathbf{x}^T C_{ce}C_{ce}\mathbf{x} = \mathbf{x}^T C_e^T C_{ce}\mathbf{x} = (C_{ce}\mathbf{x})^T C_{ce}\mathbf{x}.$$

The scalar product $(C_{ce}\mathbf{x})^T C_{ce}\mathbf{x}$ is by (8) equal to nothing else but

$$\mathbf{x}^T C_{ce}\mathbf{x} = (C_e\mathbf{x})^T C_{ce}\mathbf{x} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

The right hand side is $= S_{XX}$ in Lecture 1. Hence, we get also in Lecture 1

$$\mathbf{x}^T C_{ce}\mathbf{y} = (C_e\mathbf{x})^T C_{ce}\mathbf{y} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = S_{xy}$$

Recall the total variation SS_{T} in the Fundamental Analysis of Variance Identity in Lecture 1. SS_{T} is by the above also a quadratic form

$$SS_{\mathrm{T}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{y}^T C_{ce} \mathbf{y}$$

A is an $n \times p$ matrix. Then $\mathbf{1}_n^T A$ is $1 \times p$ vector given by the rule (1) as

$$\mathbf{1}_n^T A = \left(\sum_{j=1}^n a_{j1}, \sum_{j=1}^n a_{j2} \dots \sum_{j=1}^n a_{jp} \right)$$

which contains the column sums of A. The matrix $A\mathbf{1}_p$ is $n \times 1$ and

$$A\mathbf{1}_p = \begin{pmatrix} \sum_{j=1}^n a_{1j} \\ \sum_{j=1}^n a_{2j} \\ \vdots \\ \sum_{j=1}^n a_{nj} \end{pmatrix}.$$

contains the row sums of A.

Orthogonal Matrix

Definition

An $n \times n$ matrix A is called orthogonal, if it holds that

$$A^T A = A A^T = \mathbb{I}_n$$

This means that the column vectors in A are orthogonal.

PART 1: Mean vector, Covariance matrix, Rules of Transformation

Vector Notation: Random Vector

A random vector $\mathbf{X}$ is a column vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = (X_1, X_2, \ldots, X_n)^T$$

Each X_i is a random variable.

A Vector of sample values

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1, x_2, \ldots, x_n)^T$$

We have now x_i as a notation for an outcome of X_i[1] and $\mathbf{x}$ as an outcome of $\mathbf{X}$. Marginal cdf (=cumulative distribution function) $F_{X_i}(x_i) = P(X_i \leq x_i)$. Of course, $\mathbf{x}$ designates also a generic vector used for mathematical computation.

[1] MVP makes no distinction here. $\mathbf{x}$ is sometimes a sample vector, sometimes a random variable.

Joint cdf, Joint pdf

The joint cdf (=cumulative distribution function) of a continuous random vector $\mathbf{X}$ is

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(\mathbf{X} \leq \mathbf{x}) =$$

$$= P(X_1 \leq x_1,\ldots,X_n \leq x_n)$$

Joint probability density function (PDF)

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$$

Mean Vector

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}_{\mathbf{X}} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix} \in \mathbb{R}^n,$$

a column vector of means (=expectations) of $\mathbf{X}$, $\mu_i = E[X_i]$.

Matrix, Scalar Product

If $\mathbf{X}^T$ is the transposed column vector (i.e., a row vector), then, by (2)

$$\mathbf{X}\mathbf{X}^T = \begin{pmatrix} X_1^2 & X_1X_2 & \cdots & X_1X_n \\ X_2X_1 & X_2^2 & \cdots & X_2X_n \\ \vdots & \vdots & \ddots & \vdots \\ X_nX_1 & X_nX_2 & \cdots & X_n^2 \end{pmatrix} \tag{9}$$

is an $n \times n$ random matrix, and the scalar product,

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} X_i^2$$

is a real valued r.v..

Matrix, Scalar Product

By (9) the random matrix $(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T$ is

$$\begin{pmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_n-\mu_n) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n-\mu_n)(X_1-\mu_1) & (X_n-\mu_n)(X_2-\mu_2) & \cdots & (X_n-\mu_n)^2 \end{pmatrix}$$

Covariance Matrix of A Random Vector

Covariance matrix (also denoted by $C_{\mathbf{X}}$)

$$C := E\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\right]$$

where the array in position (i, j) is

$$c_{ij} = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

is the covariance of X_i and X_j. The variances of the components of $\mathbf{X}$ are the elements on the main diagonal, i.e.,

$$c_{ii} = E\left[(X_i - \mu_i)^2\right] = \mathrm{Var}(X_i) = \sigma_i^2.$$

Covariance & Non-linear Dependence

X and Y are independent $\Rightarrow \mathbf{Cov}(X, Y) = 0$.

The converse implication is not true in general, as shown in the next example.
Let $X \sim N(0, 1)$, the pdf of $N(0, 1)$ is denoted by $\phi(x)$. Set $Y = X^2$. Then Y is clearly functionally dependent on X. But we have

$$\mathbf{Cov}(X, Y) = E[(X \cdot Y)] - E[X] \cdot E[Y] = E\left[X^3\right] - 0 \cdot E[Y]$$

$$= E\left[X^3\right] = 0.$$

The last equality holds, since one has $g(x) = x^3\phi(x)$, so that $g(-x) = -g(x)$. Hence $E\left[X^3\right] = \int_{-\infty}^{+\infty} g(x)dx = 0$, c.f., (15) in the sequel, too.

Properties of a Covariance Matrix

- Covariance matrix is nonnegative definite, i.e., for all $\mathbf{x}$ we have

$$\mathbf{x}^T C \mathbf{x} \geq 0$$

  Hence

$$\det C \geq 0.$$

- The covariance matrix is symmetric

$$C = C^T$$

- It can be shown: every symmetric nonnegative definite matrix is a covariance matrix (for some random vector). See later.

Properties of a Covariance Matrix

The covariance matrix is symmetric

$$C = C^T$$

since

$$c_{ij} = E\left[(X_i - \mu_i)\,(X_j - \mu_j)\right]$$
$$= E\left[(X_j - \mu_j)\,(X_i - \mu_i)\right] = c_{ji}$$

Properties of a Covariance Matrix

A covariance matrix is positive definite, if

$$\mathbf{x}^T C \mathbf{x} > 0$$

holds for all $\mathbf{x} \neq \mathbf{0}$. Then

$$\det C > 0$$

(i.e. C is invertible).

Properties of a Covariance Matrix

Proposition

$$\mathbf{x}^T C\mathbf{x} \geq 0$$

Proof:

$$\mathbf{x}^T C\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j c_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

$$= \mathbf{x}^T E\left[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T\right]\mathbf{x}$$

$$= E\left[\mathbf{x}^T(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T\mathbf{x}\right] = E\left[\mathbf{x}^T\mathbf{w}\cdot\mathbf{w}^T\mathbf{x}\right]$$

where we have set $\mathbf{w} = (\mathbf{X} - \mu_{\mathbf{X}})$. Then by $\mathbf{x}^T\mathbf{w} = \mathbf{w}^T\mathbf{x} = \sum_{i=1}^{n} w_i x_i$,

$$E\left[\mathbf{x}^T\mathbf{w}\mathbf{w}^T\mathbf{x}\right] = E\left[\left(\sum_{i=1}^{n} w_i x_i\right)^2\right] \geq 0.$$

Properties of a Covariance Matrix

In terms of the entries $c_{i,j}$ of a covariance matrix $C = \left(c_{ij}\right)_{i=1,j=1}^{n,n,}$ there are the following necessary properties.

1. $c_{ij} = c_{ji}$ (symmetry).
2. $c_{ii} = \mathbf{Var}\left(X_i\right) = \sigma_i^2 \geq 0$ (the elements in the main diagonal are the variances, and thus all elements in the main diagonal are nonnegative).
3. $c_{ij}^2 \leq c_{ii} \cdot c_{jj}$ (Cauchy-Schwartz' inequality).

Coefficient of Correlation

The *Coefficient of Correlation* ρ of X and Y is defined as

$$\rho := \rho_{X,Y} := \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{Var}(X) \cdot \mathbf{Var}(Y)}},$$

where $\mathbf{Cov}(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$. This is normalized

$$-1 \leq \rho_{X,Y} \leq 1$$

For random variables X and Y,

- $\mathbf{Cov}(X, Y) = \rho_{X,Y} = 0$ does not always mean that X, Y are independent.
- $\rho_{X,Y} = \rho_{Y,X}$!

Special case: Covariance Matrix of a Bivariate Vector

$\mathbf{X} = (X_1, X_2)^T$.

$$C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where ρ is the coefficient of correlation of X_1 and X_2, and $\sigma_1^2 = \mathbf{Var}(X_1)$, $\sigma_2^2 = \mathbf{Var}(X_2)$. C is invertible iff $\rho^2 \neq 1$, to see this we note that

$$\det C = \sigma_1^2\sigma_2^2\left(1 - \rho^2\right)$$

An Example of a Covariance Matrix

Example

$\mathbb{I}_n$ is the $n \times n$ identity matrix, see (4). $\mathbb{I}_n$ is a symmetric and positive definite matrix, hence

$$\mathbb{I}_n \text{ is a (diagonal) covariance matrix.} \tag{10}$$

It is the covariance matrix of $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$, where the component r.v.'s are pairwise uncorrelated and $\mathrm{Var}[X_i] = 1$ for every i.

Simple Linear Regression Model in Matrix Form

The simple linear regression model equations for the training set

$$Y_i = \mu_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

are usefully written in matrix terms. Set

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$n \times 1$ -vector, $n \times 2$ - matrix, 2×1 -vector. The simple linear regression is now given as

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is $n \times 1$ random vector.

The Assumptions of the Simple Linear Regression Model

$$\mathbf{Y} = X\beta + \varepsilon.$$

1) **Correct**: $E[\varepsilon] = \mathbf{0}_n$ (= the $n \times 1$ zero vector), i.e., $\mathbf{0}_n \in \mathbb{R}^n$.
2) **Uncorrelated** The covariance matrix C_ε of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is

$$C_\varepsilon = \sigma^2 \mathbb{I}_n.$$

   The identity matrix $\mathbb{I}_n$ as defined (4).
3) **Homoscedastic** ε: σ^2 does not depend on X

Colored Noise

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

1) **Correct**: $E[\boldsymbol{\varepsilon}] = \mathbf{0}_n$ (= the $n \times 1$ zero vector), i.e., $\mathbf{0}_n \in \mathbb{R}^n$.
2) **Correlated** The covariance matrix $C_{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is

$$C_{\boldsymbol{\varepsilon}} = \mathbb{D}_n.$$

   where $\mathbb{D}_n$ is a positive definite matrix.
3) **Homoscedastic** Variance of ε does not depend on X

The Assumptions of the Simple Linear Regression Model

$$\mathbf{y} = X\beta + \varepsilon.$$

4) As discussed in the first lecture, 1) X can be a designed matrix, and is then not a r.v.. Or, 2) X contains the observed values $(x_1, \ldots, x_n)$ of a covariate/predicting variable. This means in particular that there is no measurement error in the covariate x. Theoretical regression line represents in general an approximation $E[\mathbf{Y}|\mathbf{X} = (x_1, \ldots, x_n)^T]$.

Computation rules

Proposition

Y *and* **X** *are random vectors,* $\mu_{\mathbf{Y}} = E[\mathbf{Y}]$, $\mu_{\mathbf{X}} = E[\mathbf{X}]$, **X** *has covariance matrix* $C_{\mathbf{X}}$, A *and* B *are* $m \times n$ *matrices.* **a** *and* **b** *are vectors of suitable dimensions. Then we have*

- $$E[\mathbf{X} + \mathbf{Y}] = \mu_{\mathbf{X}} + \mu_{\mathbf{Y}} \tag{11}$$
- $\mathbf{Z} = A\mathbf{X} + \mathbf{b}$,
$$E[\mathbf{Z}] = A\mu_{\mathbf{X}} + \mathbf{b}, \tag{12}$$
$$C_{\mathbf{Z}} = AC_{\mathbf{X}}A^T. \tag{13}$$
- $C_{\mathbf{X}} = E\left[\mathbf{X}\mathbf{X}^T\right] - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^T$
- $\mathrm{Var}\left[\mathbf{a}^T\mathbf{X}\right] = \mathbf{a}^T C_{\mathbf{X}}\mathbf{a}$

Variance Operator of MVP

The rule (13) above, i.e.

$$C_{\mathbf{Z}} = AC_{\mathbf{X}}A^T.$$

is in MVP, p. 580, 4. of C.2.3, written as

$$\mathrm{Var}\,(\mathbf{Z}) = AC_{\mathbf{X}}A^T,$$

where $\mathrm{Var}\,(\mathbf{Z})$ is called a variance operator. This notion is not found in the INDEX of MVP, and it is perhaps not really defined in the text, c.f., p. 80.

We control one of the rules above

$C = E\left[\mathbf{X}\mathbf{X}^T\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T$. By definition and matrix rules

$$C = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right] = E\left[\mathbf{X}\mathbf{X}^T - \mathbf{X}\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{X}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\right]$$

(Note that all matrices in the right hand side are $n \times n$, and hence matrix addition is defined.) Now use the rules (11) and (12)

$$= E\left[\mathbf{X}\mathbf{X}^T\right] - \underbrace{E[\mathbf{X}]}_{=\boldsymbol{\mu}}\boldsymbol{\mu}^T - \boldsymbol{\mu}\underbrace{E\left[\mathbf{X}^T\right]}_{=\boldsymbol{\mu}^T} + \boldsymbol{\mu}\boldsymbol{\mu}^T$$

$$= E\left[\mathbf{X}\mathbf{X}^T\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

This is clearly a matrix version of the univariate formula:

$$E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y.$$

Computation rules

Proposition

$\mathbf{X}$ *is a random vector,* $\boldsymbol{\mu}_{\mathbf{X}} = E[\mathbf{X}]$*, with the covariance matrix* $C_{\mathbf{X}}$*, For* A *is* $n \times n$ *matrix. Then we have*

- $E\left[\mathbf{X}^T A \mathbf{X}\right] = \operatorname{tr}(AC_{\mathbf{X}}) + \mu_{\mathbf{X}}^T A \boldsymbol{\mu}_{\mathbf{X}}$*, where* $\operatorname{tr}(B) = \sum_{i=1}^{n} b_{ii}$ *(=trace of* B*) is the sum of the entries on the main diagonal of a square matrix* B*.*

Computation rules: Proof of $C_{\mathbf{Z}} = AC_{\mathbf{X}}A^T$, where $\mathbf{Z} = A\mathbf{X} + \mathbf{b}$

By definition, $C_{\mathbf{Z}} = E\left[(\mathbf{Z} - E[\mathbf{Z}])(\mathbf{Z} - E[\mathbf{Z}])^T\right]$. By (2), $E[\mathbf{Z}] = A\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{b}$, and thus

$$\mathbf{Z} - E[\mathbf{Z}] = A\mathbf{X} + \mathbf{b} - (A\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{b}) = A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}).$$

This gives

$$C_{\mathbf{Z}} = E\left[A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}))^T\right] = E\left[A(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T A^T\right]$$

$$= AE\left[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\right] A^T = AC_{\mathbf{X}}A^T.$$

□

Diagonalizing Covariance Matrices

If C is a covariance matrix, then there exists an orthogonal matrix P such that

$$P^T C P = D, \tag{14}$$

where D is diagonal (with the eigenvalues of C on the main diagonal[2]). Suppose $\mathbf{X}$ has mean $\mathbf{0}$ and covariance matrix C. Then if $\mathbf{Y} = P^T \mathbf{X}$, $\mathbf{Y}$ has mean $\mathbf{0}$ and covariance matrix D, i.e., the components of $\mathbf{Y}$ have zero means and are pairwise non-correlated.

[2] The eigenvalues of a covariance matrix are non-negative, see Appendix

Remark: Every symmetric and nonnegative definite square matrix is covariance matrix

Let Σ be an $n \times n$ symmetric and positive definite matrix. Then Σ can be Cholesky factored to a lower triangular matrix A such that

$$\Sigma = AA^T.$$

Take a random vector $\mathbf{X}$ with $\mathbf{I}_n$ as covariance matrix. Set $\mathbf{Y} = A\mathbf{X}$. Rule (13) for computation of covariance matrices of linear transformations yields that

$$C_{\mathbf{Y}} = AC_{\mathbf{X}}A^T = A\mathbb{I}_n A^T = AA^T = \Sigma.$$

A matrix is a Cholesky factor for a covariance matrix if and only if it is lower triangular, the diagonal entries are positive,

PART 2: THE MULTIVARIATE NORMAL DISTRIBUTION

We recall first some of the properties of univariate normal distribution. Most of the facts on multivariate normal distribution stated below are found with proofs in chapter 5 of Gut, Allan: *An Intermediate Course in Probability. Second Edition Springer, 2009.*[3]

[3] You do not need these proofs in Gut loc.cit to pass this course. It is only the statements that count. But there are other proofs needed.

Normal (Gaussian) One-dimensional RVs

- X is a normal random variable if its pdf is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

  where μ is real and $\sigma > 0$.
- A symbolic shorthand notation: $X \sim N(\mu, \sigma^2)$
- Properties: $E(X) = \mu$, $\mathbf{Var}(X) = \sigma^2$

Central Moments Normal (Gaussian) One-dimensional RVs

$X \sim N(0, \sigma^2)$. Then

$$E\left[X^n\right] = \begin{cases} 0 & n \text{ is odd} \\ \frac{(2k)!}{2^k k!}\sigma^{2k} & n = 2k, k = 0, 1, 2, \ldots. \end{cases} \tag{15}$$

Linear Transformation

- $X \sim N(\mu_X, \sigma^2) \Rightarrow Y = aX + b \sim N(a\mu_X + b, a^2\sigma^2)$
- Thus $Z = \frac{X-\mu_X}{\sigma_X} \sim N(0,1)$ and

$$\mathrm{P}(X \leq x) = \mathrm{P}\left(\frac{X-\mu_X}{\sigma_X} \leq \frac{x-\mu_X}{\sigma_X}\right)$$

or

$$F_X(x) = \mathrm{P}\left(Z \leq \frac{x-\mu_X}{\sigma_X}\right) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right)$$

Multivariate Normal: Definition

Definition

An $n \times 1$ random vector $\mathbf{X}$ has a (multivariate) normal distribution iff for **every** *$n \times 1$-vector $\mathbf{a}$ the one-dimensional random vector $a^T\mathbf{X}$ has a normal distribution.*

We write $\mathbf{X} \sim N_n(\mu, C)$, when μ is the mean vector and C is the covariance matrix. Mean and covariance matrix, when they exist, do not in general determine the joint distribution of a multivariate r.v.. However, a normal random vector $\mathbf{X}$ is completely determined by μ and C.

Properties

An $n \times 1$ vector $\mathbf{X} \sim N_n(\boldsymbol{\mu}, C)$ iff the one-dimensional random variable $\mathbf{a}^T\mathbf{X}$ has a normal distribution for **every** n-vector $\mathbf{a}$.

Now we know that (take $A = \mathbf{a}^T$ and note the transformation rules above)

$$E\left[a^T\mathbf{X}\right] = \mathbf{a}^T\boldsymbol{\mu}, \mathrm{Var}\left[\mathbf{a}^T\mathbf{X}\right] = \mathbf{a}^T C\mathbf{a}$$

Hence

$$a^T\mathbf{X} \sim N\left(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T C\mathbf{a}\right)$$

Properties

Let D be a diagonal covariance matrix with σ_i^2s on the main diagonal, i.e.,

$$D = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ 0 & \ddots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{pmatrix},$$

Proposition

If $\mathbf{X} \sim N_n(\mu, D)$, then $X_1, X_2, \dots, X_n$ are independent real valued normal variables, $X_i \sim N(\mu_i, \sigma_i^2)$.

Further properties of the multivariate normal

$\mathbf{X} \sim N_n(\boldsymbol{\mu}, C)$

- Every component X_k is one-dimensional normal. To prove this we take the orthonormal basis vector

$$\mathfrak{E}_k = (0, 0, \ldots, \underbrace{1}_{\text{position } k}, 0, \ldots, 0)^T$$

  and the conclusion follows by Def. I.
- $X_1 + X_2 + \cdots X_n$ is one-dimensional normal. Note: The terms in the sum need not be independent.

Properties of the multivariate normal

$\mathbf{X} \sim N_n(\boldsymbol{\mu}, C)$

- Every marginal distribution of k variables ($1 \leq k < n$) is normal. To prove this we consider any k variables $X_{i_1}, X_{i_2} \ldots X_{i_k}$ and then take $\mathbf{a}$ such that $a_j = 0$ for $j \neq i_1, \ldots i_k$ and then apply Def. I.

Properties of multivariate normal

Proposition

$\mathbf{X} \sim N_n(\boldsymbol{\mu}, C)$ *and* $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$, *B is* $m \times n$. *Then*

$$\mathbf{Y} \sim N_m\left(B\boldsymbol{\mu} + \mathbf{b}, BCB^T\right). \tag{16}$$

Diagonalizing Covariance Matrix and Multivaraite Normals

If $\mathbf{X} \sim N_n(\mathbf{0}, C)$ and P is as in (14). Then if $\mathbf{Y} = P^T\mathbf{X}$, we have that

$$\mathbf{Y} \sim N_n(\mathbf{0}, D).$$

In other words, $\mathbf{Y}$ is a normal vector that has independent components $\sim N(0, \lambda_j)$. This trick has several important applications, e.g., for heteroscedastic multiple linear regression.

Multivariate normal: joint PDF

Definition

A random vector $\mathbf{X}$ *with mean vector* $\boldsymbol{\mu}$ *and an invertible covariance matrix* C *is* $N_n(\boldsymbol{\mu}, C)$*, iff its joint pdf is*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(C)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T C^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{17}$$

It can be shown that this definition and the definition given first (which does not require an invertible covariance matrix) are equivalent, as soon as the covariance matrix is invertible.

Multivariate normal

It can be checked that

$$\int_{R^n} \frac{1}{(2\pi)^{n/2}\sqrt{\det(C)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T C^{-1}(\mathbf{x}-\boldsymbol{\mu})} d\mathbf{x} = 1.$$

For aid on this, see EXQ in one of the Appendices below.

Standard Normal Vector: definition

$\mathbb{I}_n$ is the $n \times n$ is a covariance matrix, as pointed out in (10). Then $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbb{I}_n)$ is called a standard normal vector. When we insert $C = \mathbb{I}_n$ in (17) above we get, as $\mathbb{I}_n^{-1} = \mathbb{I}_n$,

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathbb{I}_n)}} e^{-\frac{1}{2}(\mathbf{z}-\mathbf{0})^T \mathbb{I}_n^{-1}(\mathbf{z}-\mathbf{0})} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}}$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n z_i^2} = \prod_{i=1}^n \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}z_i^2}$$

which is a product of n pdf's for i.i.d. $Z_i \sim N(0,1)$.

Rule of transformation of a pdf

Proposition

If $\mathbf{X}$ *has the joint density* $f_{\mathbf{X}}(\mathbf{x})$, $Y = A\mathbf{X} + \mathbf{b}$, *and* A *is invertible, then*

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det A|} f_{\mathbf{X}}\left(A^{-1}(\mathbf{y} - \mathbf{b})\right) \tag{18}$$

This is an example of the rule for transformation of variables in a pdf, see Thm 2.1., p. 21 in the book by Allan Gut, mentioned in one of the first slides. A proof, is necessary here, is sketched in an Appendix.

Multivariate normal and Simple Linear Refression

Simple Normal Linear Regression

Recall

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

and

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{19}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. We add a fourth assumption to the three model assumptions above: $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$. Then (19) is simple normal linear regression model in matrix form.

Simple Normal Linear Regression: Joint density of ε

Let $\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$. This we can represent ε as $\varepsilon \stackrel{d}{=} \sigma \mathbf{Z}$, $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbb{I}_n)$. Take $A = \sigma \mathbb{I}_n$ and $\mathbf{X} = \mathbf{Z}$ and $Y = \varepsilon$. Then (18) yields

$$f_\varepsilon(\mathbf{e}) = \frac{1}{|\det \sigma \mathbb{I}_n|} f_{\mathbf{Z}}\left(\sigma^{-1}\mathbf{e}\right) = \frac{1}{(2\pi)^{n/2} |\det \sigma \mathbb{I}_n|} e^{-\frac{1}{2\sigma^2}\mathbf{e}^T\mathbf{e}}$$

Here $|\det \sigma \mathbb{I}_n| = \sigma^n$, a known rule for the determinant of a diagonal matrix. Hence one gets

$$f_\varepsilon(\mathbf{e}) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\mathbf{e}^T\mathbf{e}} \tag{20}$$

Simple Normal Linear Regression: Joint density of Y

We apply (18) with $\mathbf{Y} = X\beta + \varepsilon$, f_ε, that is, $\mathbf{X} = \varepsilon$, $A = \mathbb{I}_n$ and $\mathbf{b} = X\theta$, $\det A = \det \mathbb{I}_n = 1$. Then equation (18) gives,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{|\det \mathbb{I}_n|} f_\varepsilon(\mathbf{y} - X\beta) \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}, \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-X\beta\|^2} \end{aligned} \tag{21}$$

When this is compared with the pdf in (17), one sees that

$$\mathbf{Y} \sim N_n\left(X\beta, \sigma^2 \mathbb{I}_n\right),$$

which, of course, agrees with (16) .

Simple Normal Linear Regression: Joint density of $\mathbf{Y}$

$$\mathbf{Y} \sim N_n\left(X\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n\right) \tag{22}$$

Remark

As far as this lecturer can see, (22) is nowhere to be seen in MVP. Of course, MVP does not maintain multivariate the normal distribution except in the case $n = 2$.
The statement in (22) and its derivation hold even for multiple linear regression,where X is $n \times k + 1$ (here $k = 1$) and $\boldsymbol{\beta}$ is $k + 1 \times 1$.

Maximum Likelihood Estimate (MLE) of β is LSE of β

Next the least squares criterion in simple linear regression is equal to (recall $\mathbf{x}^T\mathbf{x} = \| \mathbf{x} \|^2 = \sum_{i=1}^n x_i^2$).

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \| \mathbf{y} - X\boldsymbol{\beta} \|^2$$

Hence (21) with $k = 1$ defines the **likelihood function**

$$L_{\mathbf{y}}(\boldsymbol{\beta}) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2} Q(\beta_0, \beta_1)} \tag{23}$$

For $k = 1$, minimization of $Q(\beta_0, \beta_1)$ equivalent to maximization of $L_{\mathbf{y}}(\boldsymbol{\beta})$ (Here one is to estimate σ^2, but this does not change the MLE of $\boldsymbol{\beta}$).

Part 3: Multivariate Normal: the Bivariate Case

Multivariate Normal: the bivariate case

As soon as $\rho^2 \neq 1$, the matrix

$$C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

is invertible, and the inverse is

$$C^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

Multivariate Normal: the Bivariate Case

$\mathbf{X} = (X_1, X_2)^T$ is bivariate ($n = 2$) normal $N_2(\boldsymbol{\mu}, C)$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Assume $\rho^2 \neq 1$. Then $\mathbf{X}$ has the PDF

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det C}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T C^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}C(x_1,x_2)},$$

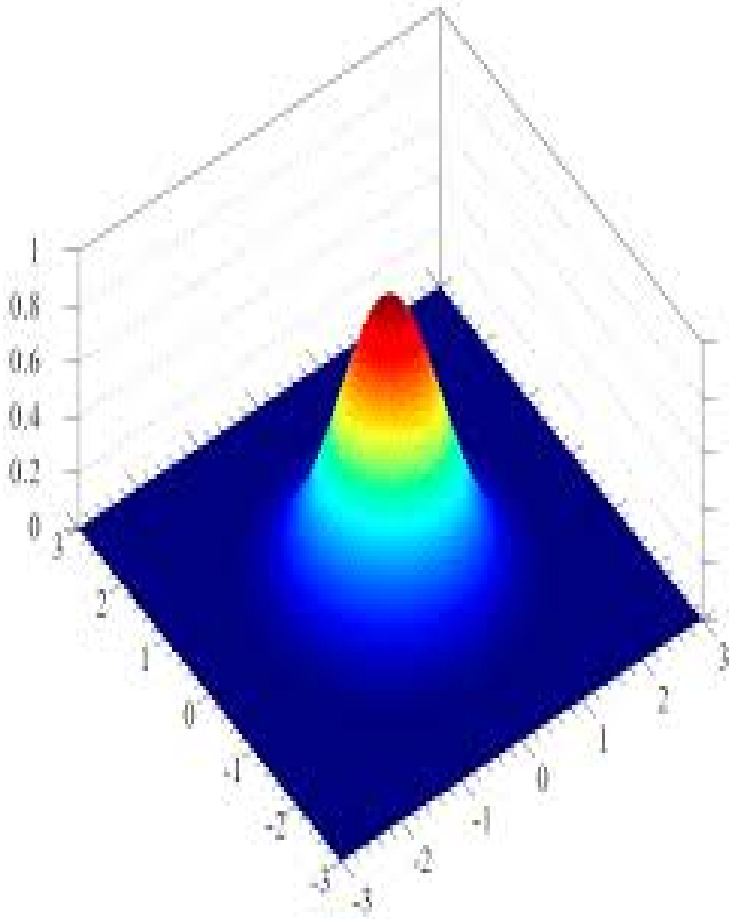
where

$$C(x_1, x_2) =$$

$$\frac{1}{(1-\rho^2)} \cdot \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right]$$

For this, invert the matrix C and expand the quadratic form !

$\rho = 0$



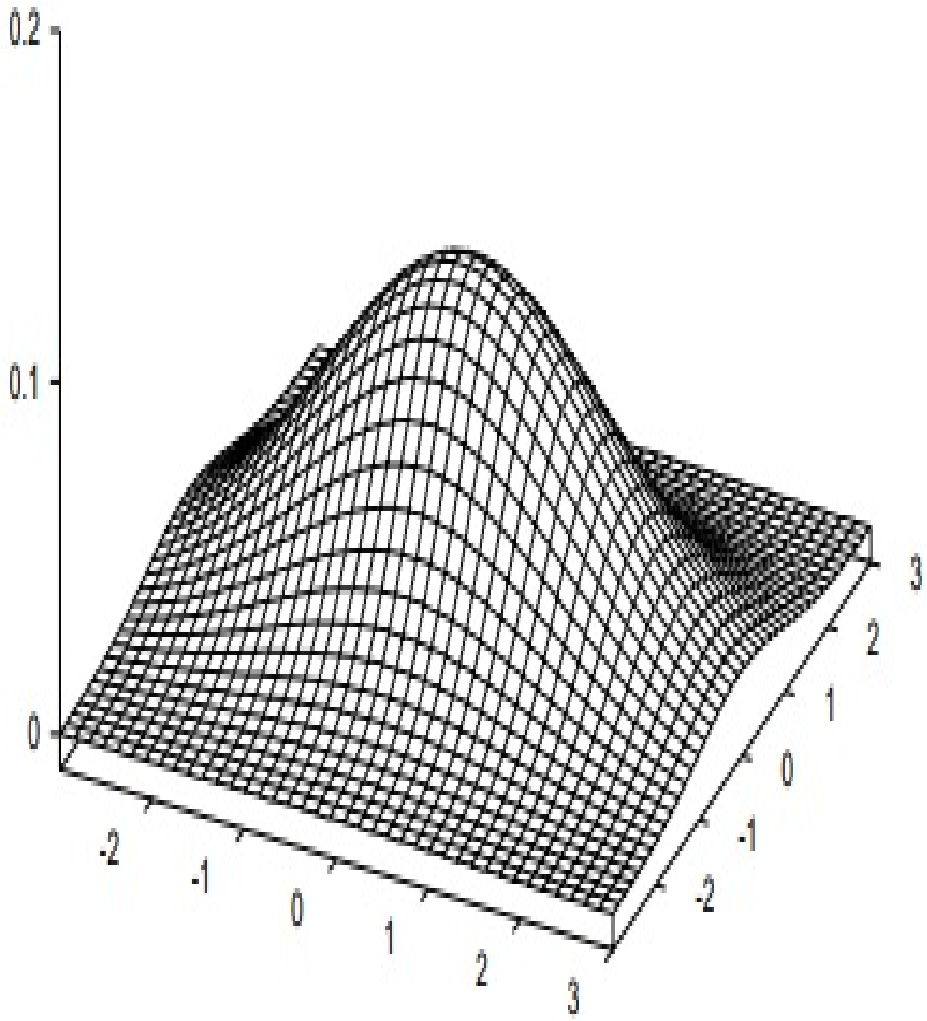


Fig. 2.2: Density functions of a two-dimensional normal distribution for uncorrelated factors, $\rho = 0$, with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1.5$, $\sigma_2 = 1.0$

Conditional densities for the bivariate normal

Complete the square of the exponent to write

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y)$$

where

$$f_X(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2}$$

$$f_{Y|X}(y) = \frac{1}{\tilde{\sigma}_2\sqrt{2\pi}} e^{-\frac{1}{2\tilde{\sigma}_2^2}(y-\tilde{\mu}_2(x))^2}$$

$$\tilde{\mu}_2(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \tilde{\sigma}_2 = \sigma_2\sqrt{1-\rho^2}$$

Bivariate normal properties

1. $E(X) = \mu_1$

2. *Given $X = x$, Y has a univariate normal distribution*

3. *Conditional mean of Y given $X = x$:*

$$\tilde{\mu}_2(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) = E(Y|X = x)$$

4. *Conditional variance of Y given $X = x$:*

$$\mathbf{Var}(Y|X = x) = \sigma_2^2\left(1 - \rho^2\right)$$

Bivariate normal properties

1. Conditional mean of Y given $X = x$:

$$\tilde{\mu}_2(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) = E(Y|X = x)$$

Bivariate normal properties

1. Conditional mean of Y given $X = x$:

$$\tilde{\mu}_2(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) = E(Y|X = x)$$

2. Conditional variance of Y given $X = x$:

$$\mathrm{Var}(Y|X = x) = \sigma_2^2\left(1 - \rho^2\right)$$

Bivariate normal properties

1. Conditional mean of Y given $X = x$:

$$\tilde{\mu}_2(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) = E(Y|X = x)$$

2. Conditional variance of Y given $X = x$:

$$\mathrm{Var}(Y|X = x) = \sigma_2^2\left(1 - \rho^2\right)$$

I.e., the conditional mean of Y given X in a bivariate normal distribution is also in the mean square sense the best LINEAR predictor of Y based on X, and the conditional variance is the variance of the estimation error. C.f., *the theoretical simple regression line in Lecture 1.*

3.

Regression in bivariate normal r.v.s is symmetric

Of course, the conditional mean of X given Y in a bivariate normal distribution is also a linear predictor, the preceding computation can be exactly repeated with $f_{X|Y}$. You can just as well predict Y by X as X by Y.

A Quote. Gorroochurn, P.: On Galton's change from "reversion" to "regression", vol. 70, 3, 227–231, The American Statistician, 2016

irrefutable confirmation needed a proper mathematical analysis, a task that Galton thought was beyond his analytical skills. Therefore, he solicited the help of the able mathematician J. Hamilton Dickson. In modern mathematical language (this is

A Quote. Gorroochurn, P.: On Galton's change from "reversion" to "regression", vol. 70, 3, 227–231, The American Statistician, 2016

face of frequency of p' (Galton 1886), Dickson was provided with the information that $Y \sim N(0, \sigma_Y^2)$ and $X|Y \sim N(\beta_{X|Y} y, \sigma_{X|Y}^2)$, and was asked the following questions:

1. What is the joint density of (X, Y), and what is the shape of the contours of equal probability density?
2. How can the regression coefficient $\beta_{Y|X}$ be calculated?
3. What is the density of Y given X?
4. What is the relationship between $\beta_{Y|X}$ and $\beta_{X|Y}$?

Dickson answered each of the above questions without much trouble, and the solution was published as an Appendix to Galton's (1886) paper "Family Likeliness in Stature." In modern

That is, the maths we are handling here is dated back to 1886 and is due to James Douglas Hamilton Dickson (1849–1931), a Scottish mathematician and expert in electricity... in-depth knowledge in fields of electricity and electrostatics and also a great interest in low temperature physics. (according to Wikipedia)

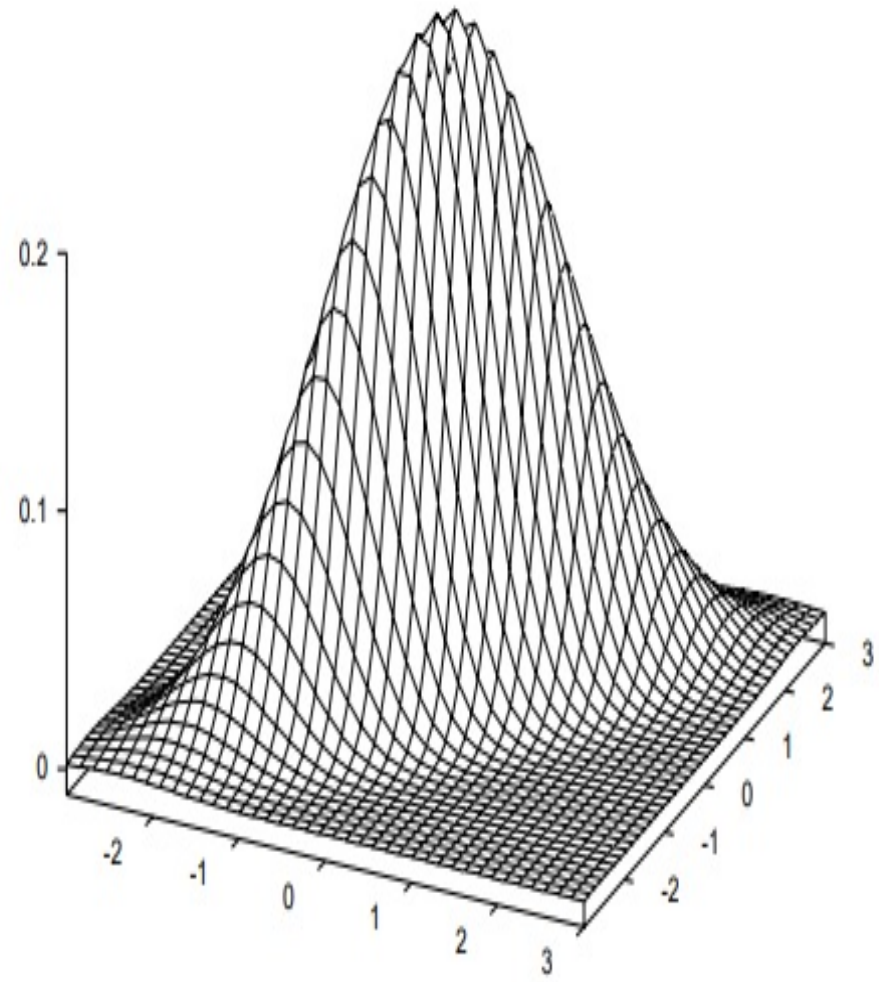


Fig. 2.3: Density functions of a two-dimensional normal distribution, $\rho = 0.8$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1.0$

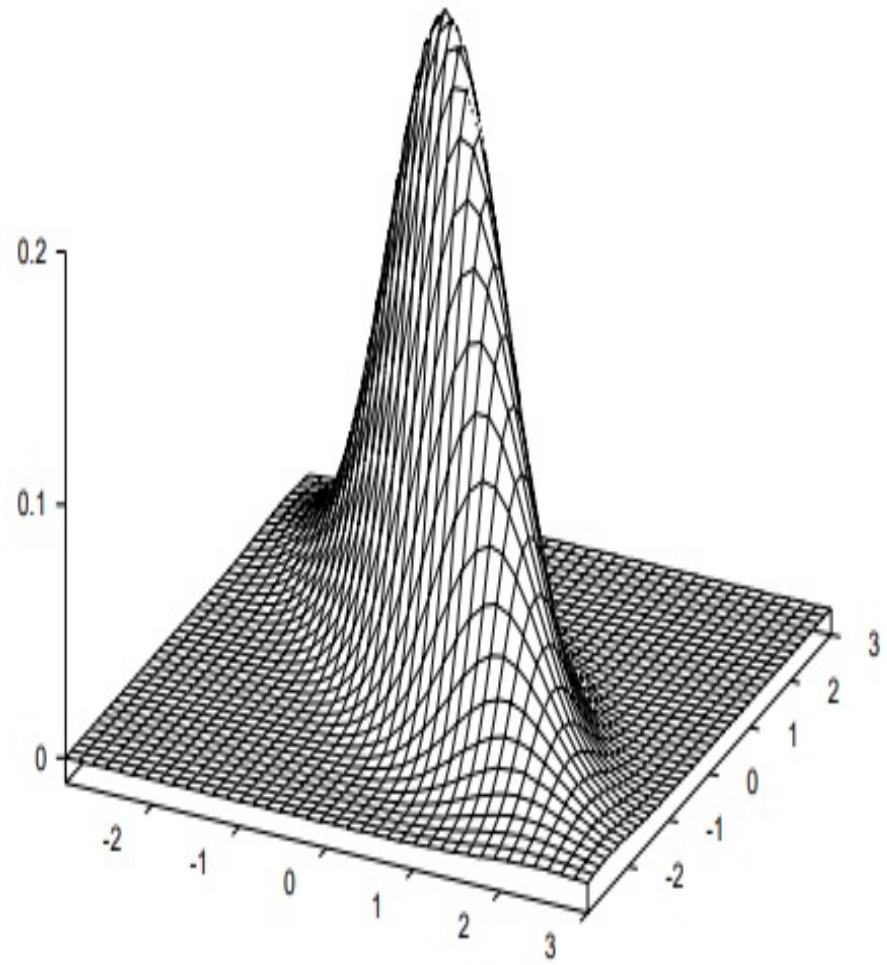


Fig. 2.4: Density functions of a two-dimensional normal distribution, $\rho = -0.8$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1.0$

Proof of conditional PDF

Consider

$$\frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{\sigma_1\sqrt{2\pi}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}Q(x,y)+\frac{1}{2\sigma_1^2}(x-\mu_1)^2}$$

where

$$-\frac{1}{2}C(x,y) + \frac{1}{2\sigma_1^2}(x-\mu_1)^2$$

$$= -\frac{1}{2}H(x,y),$$

where (next slide)

Proof of conditional PDFs

$$H(x,y) = \frac{1}{(1-\rho^2)} \cdot \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] - \left(\frac{x-\mu_1}{\sigma_1} \right)^2$$

Here we see: $\rho = 0 \Leftrightarrow$ bivariate normal X, Y are independent.
$H(x,y)$ can be rewritten as

Proof of conditional PDF

$$H(x,y) =$$

$$\frac{\rho^2}{(1-\rho^2)}\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} + \frac{(y-\mu_2)^2}{\sigma_2^2(1-\rho^2)}$$

and we get

$$H(x,y) = \frac{\left(y-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)^2}{\sigma_2^2(1-\rho^2)}$$

Conditional PDF

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{\sqrt{1-\rho^2}\sigma_2\sqrt{2\pi}} e^{\left[-\frac{1}{2}\frac{\left(y-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)^2}{\sigma_2^2(1-\rho^2)}\right]}$$

This establishes the bivariate normal properties claimed above.

Bivariate normal properties : ρ

Proposition

(X, Y) bivariate normal $\Rightarrow \rho = \rho_{X,Y}$

Proof:

$$E\left[(X-\mu_1)(Y-\mu_2)\right]$$

now use the rule of double expectation:

$$= E(E([(X-\mu_1)(Y-\mu_2)]\,|X))$$

$$= E((X-\mu_1)E\left[Y-\mu_2\right]|X))$$

Bivariate normal properties : ρ

$$= E((X-\mu_1)E\left[(Y-\mu_2)\right]|X))$$

$$= E(X-\mu_1)\left[E(Y|X)-\mu_2\right]$$

$$= E((X-\mu_1)\left[\mu_2+\rho\frac{\sigma_2}{\sigma_1}(X-\mu_1)-\mu_2\right]$$

$$= \rho\frac{\sigma_2}{\sigma_1}E\left[(X-\mu_1)(X-\mu_1)\right]$$

Bivariate normal properties : ρ

$$= \rho \frac{\sigma_2}{\sigma_1} E\left[(X-\mu_1)(X-\mu_1)\right]$$

$$= \rho \frac{\sigma_2}{\sigma_1} E\left[(X-\mu_1)^2\right]$$

$$= \rho \frac{\sigma_2}{\sigma_1} \sigma_1^2$$

$$= \rho \sigma_2 \sigma_1$$

Bivariate normal properties : ρ

In other words we have checked that

$$\rho = \frac{E\left[(X-\mu_1)(Y-\mu_2)\right]}{\sigma_2\sigma_1}$$

Appendix: Transformations, Diagonalization of Covariance Matrices

Rule of transformation of a pdf

$\mathbf{X}$ has the joint density $f_{\mathbf{X}}(\mathbf{x})$, $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, and A is invertible. Then we invert the linear (affine) transformation $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ on $\mathbb{R}^n$ by

$$\mathbf{x} = A^{-1}(\mathbf{y} - \mathbf{b})$$

Then the Jacobian matrix is

$$J = \frac{\partial}{\partial \mathbf{y}} \mathbf{x} = A^{-1}$$

and then the Jacobian determinant is

$$\det J = \frac{1}{\det A}.$$

Now we can apply Thm 2.1., p. 21 in the book by Allan Gut to obtain (18).

Standard Normal Vector by factorization of a covariance matrix

$\mathbf{X} \sim N_n(\boldsymbol{\mu}, C)$, and A is such that

$$C = AA^T$$

An invertible matrix A with this property exists always, if C is positive definite, we need the symmetry of C, too. Then

$$\mathbf{Z} = A^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

is by (16) a standard normal vector.

Factorization of a covariance matrix: the bivariate case

If

$$C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

then $C = AA^T$, where

$$A = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix},$$

Change of variables in a multiple integral

Show that

$$\int_{R^n} \frac{1}{(2\pi)^{n/2}\sqrt{\det(C)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T C^{-1}(\mathbf{x}-\boldsymbol{\mu})} d\mathbf{x} = 1.$$

Aid: It holds that $C = AA^T$. Make the change of variables $\mathbf{x} = A\mathbf{z} + \boldsymbol{\mu}$ in the multiple integral above. Then the integral can be easily evaluated. Note that the Jacobian determinant of this transformation is $\det A$. Then

$$\det C = \det A \cdot \det A^T = \det A \cdot \det A = \det A^2$$

The integrand is a pdf, hence the absolute value of the Jacobian determinant is required, so that $|\det A| = \sqrt{\det C}$.

Partioned Random Vectors and Partitioned Covariance Matrices

Let $\mathbf{X}$, $n \times 1$, be **partitioned** (= written as a vector with entries that are vectors) as

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T,$$

where $\mathbf{X}_1$ is $p \times 1$ and $\mathbf{X}_2$ is $q \times 1$, $n = q + p$. Let the covariance matrix C be **partitioned** in the sense that it is represented as a matrix with entries that are matrices, i.e.,

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{24}$$

C_{11} is $p \times p$, C_{22} is $q \times q$ e.t.c.. The mean vector is partitioned correspondingly as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \tag{25}$$

Partioned Random Vectors and Partitioned Covariance Matrices

Let $\mathbf{X} \sim N_n(\boldsymbol{\mu}, C)$, with C and $\boldsymbol{\mu}$ partitioned as in (24)-(25). Then marginal distribution of $\mathbf{X}_2$ is

$$\mathbf{X}_2 \sim N_q(\boldsymbol{\mu}_2, C_{22}).$$

as soon as C_{22} is invertible.
Then the conditional distribution of $\mathbf{X}_1$ given by $\mathbf{X}_2 = \underline{x}_2$ is

$$\mathbf{X}_1 \mid \mathbf{X}_2 = \underline{x}_2 \sim N_p\left(\underline{\boldsymbol{\mu}}_{1|2}, C_{1|2}\right),$$

where

$$\underline{\boldsymbol{\mu}}_{1|2} = \boldsymbol{\mu}_1 + C_{12}C_{22}^{-1}(\underline{x}_2 - \boldsymbol{\mu}_2)$$

and

$$C_{1|2} = C_{11} - C_{12}C_{22}^{-1}C_{21}.$$

Diagonalizable Matrices

An $n \times n$ matrix A is **orthogonally diagonalizable**, if there is an orthogonal matrix **P** such that

$$P^T A P = D,$$

where D is a diagonal matrix.

An Appendix in Matrix Calculus

Theorem

If A is an $n \times n$ matrix, then the following are equivalent:

(i) A is orthogonally diagonalizable.

(ii) A has an orthonormal set of eigenvectors.

(iii) A is symmetric.

□

Since covariance matrices are symmetric, we have by the theorem above that **all covariance matrices are orthogonally diagonalizable**.

Diagonalizable Matrices

Theorem

If A is a symmetric matrix, then

(I) Eigenvalues of A are all real numbers.

(II) Eigenvectors from different eigenspaces are orthogonal.

□

That is, **all eigenvalues of a covariance matrix are real**.

Diagonalizable Matrices

Hence we have for any covariance matrix the **spectral decomposition**

$$C = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \tag{26}$$

where $C\mathbf{u}_i = \lambda_i \mathbf{u}_i$. We normalize these and denote the orthogonal eigenvectors with $\mathbf{u}_i$. Since C is nonnegative definite,

$$0 \leq \mathbf{u}_i^T C \mathbf{u}_i = \lambda_i \mathbf{u}_i^T \mathbf{u}_i = \lambda_i,$$

and thus **the eigenvalues of a covariance matrix are nonnegative**.