

---

## TD 1: Exercises on multivariate statistics and regression

---

### ► Exercise 1

Let  $U$  and  $V$  be two independent random variables with uniform distribution over  $[0, 1]$ .

Let  $X = U + V$  and  $Y = U - V$ .

(a) Compute the expectation and covariance matrix of  $Z = \begin{pmatrix} X & Y \end{pmatrix}^T$ .

(b) Prove that  $X$  and  $Y$  are uncorrelated but not independent.

### ► Exercise 2

Let  $Z = \begin{pmatrix} X & Y \end{pmatrix}^T$  be a Gaussian vector with mean  $\mu = \begin{pmatrix} 1 & 2 \end{pmatrix}^T$  and covariance  $\Sigma = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$ .

(a) Compute the probability density function of  $Z$ .

(b) Using

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

compute the distribution of  $Y$  given  $X = x$ .

(c) What is the best prediction of  $Y$  given  $X = x$ ?

### ► Exercise 3

Consider the regression problem discussed in class: we want to determine a function  $\mu$  that takes a predictor  $X$  as input and gives the best estimate in terms of mean squared error for the observed variable  $Y$ .

In mathematical terms, we have an optimization problem defined as

$$\mu = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathbb{E}_{(X,Y)} \left[ (Y - f(X))^2 \right]$$

where  $\mathcal{F}$  is a space of functions with finite squared norm.

Show that the solution is  $\mu(x) = \mathbb{E}_{Y|X} [Y | X = x]$

### ► Exercise 4

Consider the Gaussian simple linear regression model presented in class

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad \text{with} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

The estimates for the parameters of the model,  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , are obtained  $N$  paired samples  $(x_i, y_i)$ .

(a) Show that the estimated parameters are unbiased.

(b) Show that

$$\operatorname{Var}(\hat{\beta}_1) = \frac{\sigma^2}{N} \frac{1}{s_X^2} \quad \text{and} \quad \operatorname{Var}(\hat{\beta}_0) = \frac{\sigma^2}{N} \left( 1 + \frac{\bar{X}^2}{s_X^2} \right)$$

where  $\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i$  and  $s_X^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^2$ .

Using the estimated parameters, we can predict that for a given arbitrary value of  $X$ , say  $x$  (sometimes called the operation point), we have that on average  $Y$  will be

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

(c) Show that

$$\mathbb{E}[\hat{m}(x)] = \beta_0 + \beta_1 x$$

(d) Show that the variance of  $\hat{m}(x)$  conditioned on a given choice of datapoints  $x_1, \dots, x_N$  can be written as per

$$\text{Var}_X(\hat{m}(x)) = \frac{\sigma^2}{N} \left( 1 + \frac{(x - \bar{X})^2}{s_X^2} \right)$$

Describe how the variance changes for different choices of the operation point.