

1 Gaussian vectors

1. Let Σ be a symmetric positive definite matrix of $\mathbb{R}^{n \times n}$. Provide a solution to sample a Gaussian vector with covariance matrix Σ based on i.i.d. standard Gaussian variables.

It is enough to remark that $X = \mu + \Sigma^{1/2}\varepsilon \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ and $\varepsilon \sim \mathcal{N}(0, I_d)$.

2. Let ε be a random variable in $\{-1, 1\}$ such that $\mathbb{P}(\varepsilon = 1) = 1/2$. If $(X, Y)^\top \sim \mathcal{N}(0, I_2)$ explain why the following vectors are or are not Gaussian vectors.

- (a) $(X, \varepsilon X)$.

Not Gaussian since the probability that $X + \varepsilon X = 0$ is $1/2$.

- (b) $(X, \varepsilon Y)$.

Gaussian since coordinates are independent Gaussian random variables.

- (c) $(X, \varepsilon X + Y)$.

Not Gaussian since the characteristic function of $(1 + \varepsilon)X + Y$ is not the Gaussian characteristic function.

- (d) $(X, X + \varepsilon Y)$.

Gaussian as a linear transform of (b). Indeed,

$$\begin{pmatrix} X \\ X + \varepsilon Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X \\ \varepsilon Y \end{pmatrix}.$$

3. Let X be a Gaussian vector in \mathbb{R}^n with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\sigma^2 I_n$. Prove that the random variables \bar{X}_n and $\hat{\sigma}_n^2$ defined as

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

are independent.

Let $\mathbb{1}_n$ the vector of \mathbb{R}^n with all entries equal to 1. Then, $\bar{X}_n = n^{-1} \mathbb{1}_n^\top X$ and $(n-1)\sigma_n^2 = \|X - \bar{X}_n \mathbb{1}_n\|_2^2 = \|X - n^{-1} \mathbb{1}_n \mathbb{1}_n^\top X\|_2^2 = \|(I_n - (n^{-1/2} \mathbb{1}_n)(n^{-1/2} \mathbb{1}_n)^\top)X\|_2^2$. Note that $(n^{-1/2} \mathbb{1}_n)(n^{-1/2} \mathbb{1}_n)^\top$ is the orthogonal projection onto $\text{span}(\mathbb{1}_n)$ and $I_n - (n^{-1/2} \mathbb{1}_n)(n^{-1/2} \mathbb{1}_n)^\top$ onto its orthogonal. The proof is completed by using Cochran's theorem.

2 Regression: prediction of a new observation

Consider the regression model given by

$$Y = X\beta_\star + \xi,$$

where $X \in \mathbb{R}^{n \times d}$ the $(\xi_i)_{1 \leq i \leq n}$ are i.i.d. centered Gaussian random variables with variance σ_\star^2 . Assume that $X^\top X$ has full rank and that β_\star and σ_\star^2 are estimated by

$$\hat{\beta}_n = (X^\top X)^{-1} X^\top Y \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{\|Y - X\hat{\beta}_n\|^2}{n-d}.$$

Let $x_\star \in \mathbb{R}^d$ and assume that its associated observation $Y_\star = x_\star^\top \beta_\star + \varepsilon_\star$ is predicted by $\hat{Y}_\star = x_\star^\top \hat{\beta}_n$.

1. Provide the expression of $\mathbb{E}[(\hat{Y}_\star - x_\star^\top \beta_\star)^2]$.

By definition of $\hat{\beta}_n$,

$$\hat{Y}_\star - x_\star^\top \beta_\star = x_\star^\top (\hat{\beta}_n - \beta_\star),$$

so that $\mathbb{E}[\hat{Y}_\star] = x_\star^\top \beta_\star$ and

$$\mathbb{E}[(\hat{Y}_\star - x_\star^\top \beta_\star)^2] = \mathbb{V}[\hat{Y}_\star] = x_\star^\top \mathbb{V}[\hat{\beta}_n] x_\star.$$

On the other hand,

$$\mathbb{V}[\hat{\beta}_n] = (X^\top X)^{-1} X^\top \mathbb{V}[Y] X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}.$$

Therefore,

$$\mathbb{E}[(\hat{Y}_\star - x_\star^\top \beta_\star)^2] = \sigma^2 x_\star^\top (X^\top X)^{-1} x_\star.$$

2. Provide a confidence interval for $x_\star^\top \beta_\star$ with statistical significance $1 - \alpha$ for $\alpha \in (0, 1)$.

By the first question, \hat{Y}_\star is a Gaussian random variable with mean $x_\star^\top \beta_\star$ and variance $\sigma_\star^2 x_\star^\top (X^\top X)^{-1} x_\star$. If $z_{1-\alpha/2}$ is the quantile of order $1 - \alpha/2$ of the standard Gaussian variable,

$$\mathbb{P} \left(\frac{|\hat{Y}_\star - x_\star^\top \beta_\star|}{\sigma_\star (x_\star^\top (X^\top X)^{-1} x_\star)^{1/2}} \leq z_{1-\alpha/2} \right) \geq 1 - \alpha.$$

Therefore, with probability larger than $1 - \alpha$,

$$x_\star^\top \beta_\star \in \left(\hat{Y}_\star - \sigma_\star (x_\star^\top (X^\top X)^{-1} x_\star)^{1/2} z_{1-\alpha/2}; \hat{Y}_\star + \sigma_\star (x_\star^\top (X^\top X)^{-1} x_\star)^{1/2} z_{1-\alpha/2} \right).$$

3 Regression: linear estimators

Consider the regression model given, for all $1 \leq i \leq n$, by

$$Y_i = f^\star(X_i) + \xi_i,$$

where for all $1 \leq i \leq n$, $X_i \in \mathbf{X}$, and the $(\xi_i)_{1 \leq i \leq n}$ are i.i.d. centered Gaussian random variables with variance σ^2 . In this exercise, f^\star is estimated by a linear estimator of the form

$$\hat{f}_n : x \mapsto \sum_{i=1}^n w_i(x) Y_i.$$

Prove that

$$\frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n (\hat{f}_n(X_i) - f^\star(X_i))^2 \right] = \frac{1}{n} \|W f^\star(X) - f^\star(X)\|_2^2 + \frac{\sigma^2}{n} \text{Trace}(W^\top W),$$

where $W = (w_i(X_j))_{1 \leq i, j \leq n}$ and $f^\star(X) = (f^\star(X_1), \dots, f^\star(X_n))^\top$.

Note that

$$\frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n (\hat{f}_n(X_i) - f^\star(X_i))^2 \right] = \frac{1}{n} \mathbb{E} [\|WY - f^\star(X)\|_2^2],$$

where $Y = (Y_1, \dots, Y_n)^\top$. then, write

$$\begin{aligned} \mathbb{E} [\|WY - f^\star(X)\|_2^2] &= \mathbb{E} [\|WY - Wf^\star(X)\|_2^2] + \mathbb{E} [\|Wf^\star(X) - f^\star(X)\|_2^2] \\ &\quad + 2\mathbb{E} [\langle WY - Wf^\star(X); Wf^\star(X) - f^\star(X) \rangle]. \end{aligned}$$

As $\mathbb{E}[Y] = f^*(X)$, this yields

$$\mathbb{E} [\|WY - f^*(X)\|_2^2] = \mathbb{E} [\|WY - Wf^*(X)\|_2^2] + \|Wf^*(X) - f^*(X)\|_2^2.$$

The proof is completed by noting that

$$\mathbb{E} [\|WY - Wf^*(X)\|_2^2] = \mathbb{E} [(Y - f^*(X))^\top W^\top W (Y - f^*(X))] = \text{Trace} (W^\top W \mathbb{V}[Y - f^*(X)])$$

and $\mathbb{V}[Y - f^*(X)] = \sigma^2 I_n$.