Homework 5

Solutions: 14.05.2025

1. Let X_j be *jointly Gaussian*, i.e., they are the entries of a d-dimensional Gaussian random vector $X = (X_1, \dots, X_d)$, with $X \sim \mathcal{N}(m, \Sigma)$.

   (a) Prove that the projections of the vector X on a subset $I \subset \{1, \dots, d\}$ of its coordinates is a $|I|$-dimensional Gaussian with mean vector $(m_j)_{j \in I}$, and covariance matrix $(\Sigma_{ij})_{ij \in I^2}$.
   Hint: Write the projection on a subset of the component as a matrix $\Pi \in \{0,1\}^{|I| \times d}$ and use the formula for the linear transformation of Gaussian random variables.

   (b) Prove that if, for a subset $I \subset \{1, \dots, d\}$ of the coordinates we have $\Sigma_{ij} = 0$ for all $i, j \in I, i \neq j$, then the random variables $\{X_j\}_{j \in I}$ are mutually independent.

2. Consider a 2-dimensional random vector (X_1, X_2) so that $X_1 \sim \mathcal{N}(0,1)$ and $X_2 = W X_1$ where W is a Rademacher random variable ($\mathbb{P}(W = 1) = \mathbb{P}(W = -1) = 1/2$). Prove that (X_1, X_2) is not a 2-dimensional Gaussian distribution.
   Hint: one possible solution is to show that $X_1 + X_2$ is not a 1-dimensional Gaussian.

3. Let $(X_1, \dots, X_p)$ and $(Y_1, \dots, Y_q)$ be two independent Gaussian samples with respective distributions $\mathcal{N}(m_1, \sigma^2)$ and $\mathcal{N}(m_2, \sigma^2)$. We define

$$\bar{X} = \frac{\sum_{i=1}^{p} X_i}{p}, \qquad S_X^2 = \frac{\sum_{i=1}^{p} (X_i - \bar{X})^2}{p-1},$$

$$\bar{Y} = \frac{\sum_{i=1}^{q} Y_i}{q}, \qquad S_Y^2 = \frac{\sum_{i=1}^{q} (Y_i - \bar{Y})^2}{q-1}.$$

   Show that:

   (a) The random variable $\bar{X} - \bar{Y}$ follows the distribution $\mathcal{N}(m_1 - m_2, \sigma^2 \left(\frac{1}{p} + \frac{1}{q}\right))$;

   (b) The random variable $S_X^2(p-1) + S_Y^2(q-1)$ follows the distribution $\sigma^2 \cdot \chi^2(p+q-2)$;

   (c) The random variables $\bar{X}, \bar{Y}$ and $S_X^2(p-1) + S_Y^2(q-1)$ are (jointly) independent;

   (d) The random variable

$$\frac{\sqrt{p+q-2} \cdot \left(\bar{X} - \bar{Y} - (m_1 - m_2)\right)}{\sqrt{\frac{1}{p} + \frac{1}{q}} \cdot \sqrt{\sum_{i=1}^{p} (X_i - \bar{X})^2 + \sum_{i=1}^{q} (Y_i - \bar{Y})^2}}$$

   follows the t distribution with $p + q - 2$ degrees of freedom.

4. Consider a linear model given by

$$Y = \Phi \vartheta + \sigma Z$$

   where $\Phi \in \mathbb{R}^{n \times p}$, $\vartheta \in \mathbb{R}^p$, $Z \sim \mathcal{N}(0, \mathbb{1}_n)$. Assume that the model is overparametrized, i.e., $\operatorname{rank}(\Phi) = p \leq n$. Recall that the linear estimator for ϑ seen in class is given by $\hat{\vartheta}(Y) = (\Phi^\top \Phi)^{-1} \Phi^\top Y$.

   (a) Write the (joint) density of $Y = (y_1, \dots, y_n)$ and find the Maximum Likelihood Estimator for $(\vartheta, \sigma^2) \in \mathbb{R}^{p+1}$. Compare it with the estimator for (ϑ, σ^2) found in class.

   (b) Find the distribution of $\hat{\vartheta}(Y)$, and of $\|\hat{\vartheta}(Y) - \theta\|^2$

   (c) Find the distribution of $\Phi \hat{\vartheta}(Y)$ and of $\|\Phi(\vartheta - \hat{\vartheta}(Y))\|^2$.

   (d) Prove that the random variables appearing in the previous point are independent.

(e) Which ones of the above distributions have a density with respect to Lebesgue measure?

item For $i \in \{1, \dots, n\}$, let x_i be independent random variables, with x_i uniformly distributed in $\{-1, +1\}$ and ε_i uniformly distributed on $\{-\sigma, \sigma\}$, where $\sigma \in \mathbb{R}$ is unknown. For an unknown parameter $\beta \in \mathbb{R}$, let $y_i = \beta x_i + \varepsilon_i$.

(a) Calculate the least squares estimator $\hat{\beta}$ of β.

(b) Calculate its expectation and variance.

(c) Compare fixed-design case (i.e. if the x_i are not random) and the case where the x_i are random: Does the distribution of $\hat{\beta}$ differ? What if $x_i \sim \mathcal{N}(0,1)$ and $\varepsilon_i \sim \mathcal{N}(0,1)$?

(d) If x_i and ε_i are dependent, it can be impossible to find β, even as $n \to \infty$. Can you you give an example of such a situation, expressing, e.g., x_i as a particular function of ε_i?

5. Consider a linear model where datapoints are generated by

$$y_j = \phi(x_j)^\top \vartheta + \sigma Z$$

where $\phi(x_j) \in \mathbb{R}^p$, $\vartheta \in \mathbb{R}^p$, $\mathbb{E}[Z_j] = 0$, $\mathrm{Cov}(Z_j, Z_k) = \delta_{jk}$ for all $j, k \in \mathbb{N}$. Here x_j is a sequence given in advance. Furthermore, assume that the parameter space Θ is a compact subset of $\mathbb{R}^{p+1}$, that the smallest eigenvalue of the matrix $\Phi^\top \Phi = \sum_j^n \phi(x_j)\phi(x_j)^\top$ diverges with n

(a) Compute the mean and the covariance of the vector $(\Phi^\top \Phi)^{1/2}(\hat{\vartheta} - \vartheta)$

(b) Prove that $\hat{\vartheta}$ is a consistent estimator for θ.

(c) (Optional) Assuming that $\max_j \phi(x_j)(\sum_i \phi(x_i)\phi(x_i)^\top)^{-1}\phi(x_j)^\top$ converges to 0 in n, prove asymptotic normality of the estimator $\hat{\vartheta}$.