3^a Prova - FECD B - 2025/01

Renato Assunção

1. **20 points. Sufficient statistics for the Gaussian case.** Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. Use the factorization theorem to obtain a bi-dimensional sufficient statistic $T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}))$ for the parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)$. The Gaussian density is
$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

2. **20 points. Sufficient statistics for the simple linear regression model.** Let $Y_1, \ldots, Y_n$ be independent random variables with the same variance but with mean μ_i that changes with the index i. Assume the linear regression model where $\mu_i = \beta_0 + \beta_1 x_i$. Hence,
$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2), = \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2),$$
where $x_i \in \mathbb{R}$ are fixed known values (the regressor), and $\beta_0, \beta_1 \in \mathbb{R}$, $\sigma^2 > 0$ are unknown parameters.

   In the previous problem, substitute the constant μ by this changing μ_i and manipulating the expressions to show that
$$T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}), T_3(\mathbf{Y})) = \left(\sum_i y_i, \sum_i x_i y_i, \sum_i y_i^2\right)$$
   are sufficient statistics for the parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$.

3. **20 points.** Show that the binomial distribution $\mathcal{B}(n, \theta)$ belongs to the exponential distribution class by showing that, for $k = 0, 1, \ldots, n$, we have
$$\mathbb{P}(Y = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k} = h(k)g(\theta)\exp(\eta(\theta)T(k))$$
   identifying each of the functions below (HINT: $\eta(\theta) = \log(\theta/(1-\theta)$ and $T(k) = k$).

4. **20 points.** Let $Y_1, \ldots, Y_n$ be i.i.d. random variables from an unknown distribution $g(y)$. Suppose we model the data using a parametric family $\mathcal{F} = \{f(y; \theta), \theta \in \Theta\}$, which may or may not contain the true distribution g.

   Consider the Kullback-Leibler (KL) divergence from g to a model distribution $f(y; \theta)$. Explain why minimizing the KL divergence over θ is equivalent to maximizing the expected log-likelihood $\mathbb{E}_g[\log f(Y; \theta)]$.

5. **20 points.** Considere um modelo de mistura: considera-se que os dados $y_1, \ldots, y_n$ são independents e possuem distribuição $f(y; \theta_1)$ com probabilidade π ou distribuição $f(y; \theta_2)$ com probabilidade $1 - \pi$. Os parâmetros π, θ_1, θ_2 são desconhecidos, bem como de qual das duas distaribuições cada dado proven. Descreva os passos E e M do algoritmo EM. para obter o MLE dos parâmetros desconhecidos.

6. **BONUS: 10 extra points** Explain why the MLE converges to the value θ^* that minimizes the KL divergence from g to f_θ as the sample size $n \to \infty$.

Solutions:

1. The joint density of $Y_1, \ldots, Y_n$, where each $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, is given by:
$$f(\mathbf{y}; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right).$$
   We can expand the quadratic term:
$$\sum_{i=1}^n (y_i-\mu)^2 = \sum y_i^2 - 2\mu\sum y_i + n\mu^2.$$

So the joint density becomes:

$$f(\mathbf{y};\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left[\sum y_i^2 - 2\mu\sum y_i + n\mu^2\right]\right). = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{n\mu^2}{2\sigma^2}\right)\exp\left(-\frac{\sum y_i^2}{\sigma^2} + \frac{\mu\sum y_i}{2\sigma^2}\right).$$

This factorizes as:

$$f(\mathbf{y};\mu,\sigma^2) = k(\boldsymbol{\theta})\cdot g(T_1(\mathbf{y}), T_2(\mathbf{y});\boldsymbol{\theta}),$$

where

$$T_1(\mathbf{y}) = \sum y_i, \qquad T_2(\mathbf{y}) = \sum y_i^2.$$

The sufficient statistic is:

$$T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y})) = \left(\sum_{i=1}^n Y_i,\ \sum_{i=1}^n Y_i^2\right).$$

Here, T_1 captures the sample total (from which the mean is computed), and T_2 captures the total sum of squares (from which the empirical variance is computed). This is sufficient to estimate (μ, σ^2).

2. The density of each $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, with $\mu_i = \beta_0 + \beta_1 x_i$. Therefore, the joint density is:

$$f(\mathbf{y};\beta_0,\beta_1,\sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right).$$

As in the previous problem, expand the quadratic term and substitute μ_i for $\beta_0 + \beta_1 x_i$:

$$\begin{aligned}\sum_{i=1}^n (y_i-\mu_i)^2 &= \sum_i y_i^2 - 2\sum_i \mu_i y_i + \sum_i \mu_i^2 \\ &= \sum_i y_i^2 - 2\sum_i (\beta_0+\beta_1 x_i) y_i + \sum_i (\beta_0+\beta_1 x_i)^2\end{aligned}$$

So the joint density becomes:

$$\begin{aligned}f(\mathbf{y};\mu,\sigma^2) &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left[\sum y_i^2 - 2\sum y_i\mu_i + \sum_i \mu_i^2\right]\right). \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_i(\beta_0+\beta_1 x_i)^2}{2\sigma^2}\right)\exp\left(-\frac{\sum y_i^2}{\sigma^2} + \frac{\sum y_i(\beta_0+\beta_1 x_i)}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_i(\beta_0+\beta_1 x_i)^2}{2\sigma^2}\right)\exp\left(-\frac{\sum y_i^2}{\sigma^2} - \frac{\beta_0\sum y_i}{2\sigma^2} - \frac{\beta_1\sum_i y_i x_i}{2\sigma^2}\right) \\ &= k(\boldsymbol{\theta})\cdot\exp\left(-\frac{T_3(\mathbf{y})}{\sigma^2} - \frac{\beta_0 T_1(\mathbf{y})}{2\sigma^2} - \frac{\beta_1 T_2(\mathbf{y})}{2\sigma^2}\right)\end{aligned}$$

and the sufficient statistic is:

$$T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}), T_3(\mathbf{Y})) = \left(\sum_i Y_i, \sum_i Y_i x_i, \sum_i Y_i^2\right)$$

which are sufficient for estimating $(\beta_0, \beta_1, \sigma^2)$.

3. The probability mass function is

$$\mathbb{P}(Y=k) = \binom{n}{k}\theta^k(1-\theta)^{n-k} = \binom{n}{k}\cdot\left(\frac{\theta}{1-\theta}\right)^k\cdot(1-\theta)^n = \binom{n}{k}\cdot(1-\theta)^n\cdot\exp\left(k\log\left(\frac{\theta}{1-\theta}\right)\right).$$

Therefore, we can write the PMF as:

$$\mathbb{P}(Y=k) = h(k)\cdot g(\theta)\cdot\exp\left(\eta(\theta)\cdot T(k)\right),$$

with the following identifications:

- $h(k) = \binom{n}{k}$

- $g(\theta) = (1-\theta)^n$
- $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$
- $T(k) = k$

4. The Kullback-Leibler (KL) divergence from g to a model distribution $f(y;\theta)$ is:

$$\mathrm{KL}(g \,\|\, f_\theta) = \mathbb{E}_g\left[\log \frac{g(Y)}{f(Y;\theta)}\right] = \mathbb{E}_g[\log g(Y)] - \mathbb{E}_g[\log f(Y;\theta)].$$

   As the first term does not depend on θ, minimizing the KL divergence over θ is equivalent to maximizing the second term, the expected log-likelihood $\mathbb{E}_g[\log f(Y;\theta)]$.

5. We want to understand why the maximum likelihood estimator (MLE) converges to the parameter value

$$\theta^* = \arg\min_{\theta\in\Theta} \mathrm{KL}(g \,\|\, f_\theta) = \arg\max_{\theta\in\Theta} \mathbb{E}_g[\log f(Y;\theta)].$$

   Consider the log-likelihood $\ell_n(\theta)$ based on a random sample $Y_1, \ldots, Y_n$ from g. By the Law of Large Numbers (LLN), under regularity conditions, we have:

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i;\theta). \xrightarrow{a.s.} \mathbb{E}_g[\log f(Y;\theta)] \quad \text{for each } \theta \in \Theta. \tag{1}$$

   The MLE is the maximizer of the left-hand side of (1):

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \ell_n(\theta),$$

   As a consequence, the MLE should converge to the value

$$\theta^* = \arg\max_{\theta\in\Theta} \mathbb{E}_g[\log f(Y;\theta)] = \arg\min_{\theta\in\Theta} \mathrm{KL}(g \,\|\, f_\theta),$$

   that maximizes the right-hand side of (1).