## 1 Foundations of Data Science (DJW)

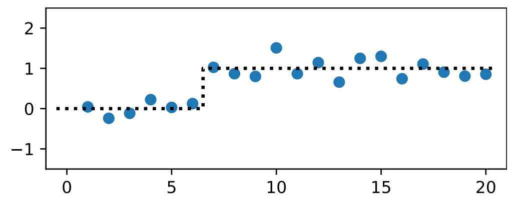
(*a*) A 0/1 signal is being transmitted. The transmitted signal at timeslot $i \in \{1, \ldots, n\}$ is $x_i \in \{0, 1\}$, and we have been told that this signal starts at 0 and then flips to 1, i.e. there is a parameter $\theta \in \{1, \ldots, n-1\}$ such that

$$x_i = \begin{cases} 0 & \text{for } i \leq \theta, \\ 1 & \text{for } i > \theta, \end{cases}$$

but the value of $\theta$ is unknown. The channel is noisy, and the received signal in timeslot $i$ is

$$Y_i \sim x_i + \text{Normal}(0, \varepsilon^2)$$

where $\varepsilon$ is known.




(*i*) Given received signals $(y_1, \ldots, y_n)$, find an expression for the log likelihood, $\log \Pr(y_1, \ldots, y_n\, ;\, \theta)$. [5 marks]

(*ii*) Give pseudocode for finding the maximum likelihood estimator $\hat{\theta}$. [3 marks]

(*b*) The Gaussian Mixture Model with $m$ components can be written as follows: first generate $K \in \{1, \ldots, m\}$, $\mathbb{P}(K = k) = p_k$, and then generate $X \sim \text{Normal}(\mu_K, \sigma_K^2)$. Here $p_1, \ldots, p_m$ and $\mu_1, \ldots, \mu_m$ and $\sigma_1, \ldots, \sigma_m$ are unknown parameters, with $p_k > 0$ and $\sigma_k > 0$ for all $k$, and $p_1 + \cdots + p_m = 1$. The number of components $m$ is known.

(*i*) Give formulae for $\mathbb{P}(X \leq x \mid K = k)$ and for $\mathbb{P}(X \leq x)$. [3 marks]

*You should leave your answers in terms of the cumulative distribution function for a Normal distribution,* $\Phi_{\mu,\sigma}(x) = \mathbb{P}\big(\text{Normal}(\mu, \sigma^2) \leq x\big)$.

(*ii*) Calculate the density of $X$. [4 marks]

(*iii*) Given a dataset $(x_1, \ldots, x_n)$, explain how to fit the unknown parameters using numerical optimization. [5 marks]