J Kohonen

4B Parameter estimation

About notation. Here we will use notation like $f_\lambda(x)$ to mean "the density function of the given form, when its **parameter** has value λ". Note that the subscript now refers to the parameter and *not* the random variable, like in $f_X(x)$. You must understand from context how the subscript is meant. — There are also other ways of showing the parameter; Ross writes $f(x \mid \lambda)$, and some people write $f(x\,;\,\lambda)$. Varying notation is a fact of life.

Class problems

4B1 (Service requests) A computing server receives service requests at random intervals. The intervals between each two consecutive requests are independent, and follow the density function

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

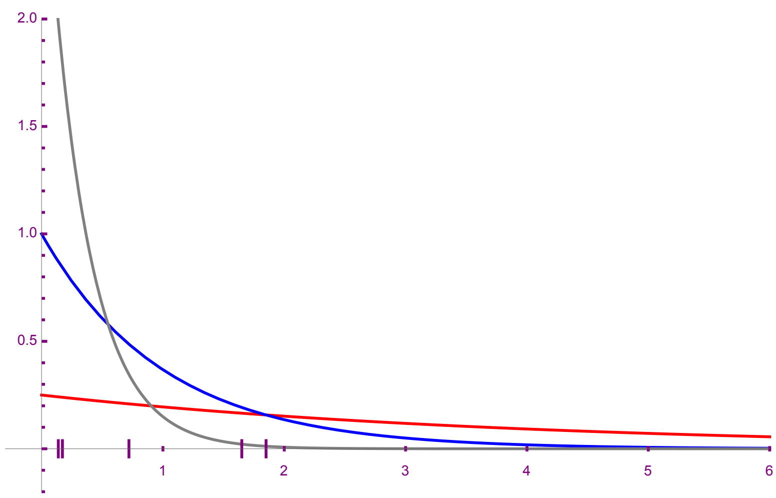
where $\lambda > 0$ is an unknown parameter. We have measured the intervals $0.16, 1.85, 0.15, 0.72, 1.65$.

(a) In the graph below, the measured values are marked on the x axis as small bars. There are also three proposed density functions for the data, corresponding to the parameter values $\lambda = 0.25$ (red), $\lambda = 1.00$ (blue) and $\lambda = 3.00$ (gray).

By *looking* at the data, give your opinion on which of the three proposed density functions might the best match for (the empirical distribution of) the data.

(b) Find the maximum likelihood estimate for the parameter λ.

The likelihood function $L(\lambda)$ and its logarithm $\ell(\lambda) = \log(L(\lambda))$ are maximized at the same value of λ, so you can use either function. The latter may have more convenient derivatives.



4B2 (Serial numbers; see Lecture 4A) A foreign army has b battle tanks (number unknown to us) with gearboxes serially numbered $1, 2, \ldots, b$. Our army has captured four tanks with serial numbers $x_1 = 13$, $x_2 = 77$, $x_3 = 111$ and $x_4 = 145$. We assume that each captured tank was independently random, with serial number uniformly distributed in the set $\{1, 2, \ldots b\}$. (Here we assume that the same number could be captured again, but hopefully this is a small error.)

(a) Having observed those numbers, is it possible that $b = 140$? What about $b = 200$? (Use common sense.)

(b) If the enemy has b tanks, and $b < 145$, what is the probability of observing this sequence of serial numbers?

(c) If the enemy has b tanks, and $b \geq 145$, what is the probability of observing this sequence of serial numbers?

(d) Write the likelihood function $L(b)$, based on the four observations, as a simple expression that is valid for all positive integers b. (Hint: Break into two cases, small and large b.)

(e) Examine the expression of $L(b)$, and find which value of b maximizes it. In other words, find the maximum likelihood estimate (MLE) for b.

(f) Now generalize. If we have seen n tanks with serial numbers $(x_1, \ldots, x_n)$, what is the MLE for b?

(g) Using the formula from the (f), if we have captured just one tank ($n = 1$) whose serial number is x_1, what is the MLE for b? Do you think it is a good estimate?

Home problems

4B3 (Continuous uniform distribution) We are observing independent random numbers from the continuous uniform distribution over the interval $[0, b]$, which has density

$$f_b(x) = \begin{cases} \frac{1}{b}, & 0 \leq x \leq b, \\ 0, & \text{elsewhere.} \end{cases}$$

The parameter b is an unknown positive real number.

(a) We have observed five numbers $(1.3, 1.9, 3.6, 5.1, 0.3)$. Write a simple expression for the likelihood function of b (hint: consider two cases). Plot the function, either by hand or by computer, over the interval $b \in [1, 10]$. Explain the shape of the function in words. This is very similar to the tank problem, except that the parameter b and the observations need not be integers.

(b) Find the ML estimate $\hat{b}$ based on these data.

(c) Generalize. If we have observed n numbers $\vec{n} = (x_1, \ldots, x_n)$, what is the MLE for b?

(d) Recall that the parameter's true value is b (but we do not know the value). Suppose we only observe one point X_1 (using uppercase to highlight that it is random, over the interval $[0, b]$). What is the expected value of X_1? What is the expected value of the ML estimator $\hat{b}$? Is the estimator biased or unbiased; if biased, up or down?

(e) Let us consider another possible estimator (different from the MLE). If the true parameter value is b, then each observation will have expected value $\mu = \mathrm{E}(X_i) = b/2$. Since $m(\vec{x})$ is a good estimator for μ, it would seem natural that $2m(\vec{x})$ would be a good estimator for $2\mu = b$. So let us define a new estimator

$$\tilde{b}(\vec{x}) = 2m(\vec{x}) = \frac{2}{n}\sum_{i=1}^{n} x_i.$$

Examine whether our new estimator $\tilde{b}(\vec{X})$ is biased or unbiased, when each observation X_i comes from the uniform distribution over $[0, b]$. (Hint: take expected value by terms.)

(f) Calculate the new estimate $\tilde{b}$ from three observations $\vec{x} = (2, 3, 16)$. Is the value reasonable?

4B4 (Fitting a geometric distribution) A random variable X has *geometric distribution* with parameter p, that is, it has density

$$f_p(x) \;=\; \begin{cases} p(1-p)^x, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise.} \end{cases}$$

Application: X is obtained when an experiment has a constant probability p of succeeding, each time; the experiment is repeated until it succeeds, and we count the number of failures. For example, tossing a coin until heads is obtained, or asking random people until you find a supporter of party P.

From this distribution, we have three independent observations $x_1 = 5$, $x_2 = 3$ and $x_3 = 10$. Find the maximum likelihood estimate for the parameter p. Looking at the value of p, explain what kind of experiment might have produced the data.

Using the logarithmic likelihood is probably convenient here.