Exercise week 3

Gherardo Varando, gherardo.varando@math.ku.dk

04/12/2019

1 A trigonometric density

Load the data in the file `angles.txt`, the data have been generated from a density $f(x|k) \propto \sin(x)^k$ in the interval $[0, \pi]$ ($f(x|k) = 0$ outside the interval $[0, \pi]$). Notation: $\propto$ stands for *proportional to* it means that $f(x|k) = C_k \sin(x)^k$, where C_k is an appropriate normalization constant.

Ex 1

Ex 1.0 How can you compute the normalization constant C_k?

EX 1.1 The model is parametric ? Which are the parameters of the model?

Ex 1.1 Write the minus log-likelihood function of the model and implement it in an R function.

Ex 1.2 Use numerical optimization method to find the maximum likelihood estimator.

Ex 1.3 Plot the histogram of the data and the density corresponding to the MLE.

2 A case study of neuronal data

We continue the case study of the ISI data in the `neuronspikes.txt` file. We are now able to estimate parameters with MLE.

Ex 2

Ex 2.1 If we assume the ISI observations are i.i.d. following an exponential distribution with parameter λ. Compute the maximum likelihood estimate of λ.

Ex 2.2 Assume now that the ISI observations are i.i.d. following a gamma distribution with parameters α (`shape`) and β (`rate`), find the MLE estimates of the parameters α and β.

Ex 2.3 For the gamma distribution we know the formulas for the mean value and the variance, as following,

$$\mathbb{E}(X) = \frac{\alpha}{\beta}$$

$$\mathbb{V}(X) = \frac{\alpha}{\beta^2}$$

Try to find the method of moments estimator of α and β. The method of moments can be used to find the first estimation to initialize the MLE iterative algorithm.

Ex 3 In addition to the exponential and the gamma distribution, the **inverse Gaussian distribution** is another widely used model for inter-events intervals. It describes the first-passage time of a one-dimensional Brownian motion subject to a fixed threshold value. The probability density function is given by,

$$f(x|\mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right)$$

Ex 3.1 Write (analytically) the formula for the log-likelihood given n i.i.d. observations.

Ex 3.2 Try to derive the formula for the maximum likelihood estimators for μ and λ (if not able skip to point 3.4)

Ex 3.3 Apply the MLE estimators in the previous step to the experimental ISI data, that is calculate the theoretical estimates of μ and λ for the ISI data.

Ex 3.4 Find the maximum likelihood estimators using numerical methods.

Ex 3.5 Plot the estimated inverse Gaussian density on top of the histogram of the ISI data and with the kernel density estimation. If you can find the method of moments estimators of the parameters, you can use those as initial points for the numerical optimization.

3 Brain cell dataset

We continue here the study of the brain cell dataset from the Allen Institute.

Ex 4

Ex 4.1 Find numerically the MLE estimates of the parameters of the log-normal distribution for the *ramp spike time* observations.

Ex 4.2 As the name suggest the log-normal distribution is related to the Gaussian distribution. In particular if X is a log-normal distribution with parameters μ and σ then $\log(X)$ is a normal distribution with mean value μ and standard deviation σ. We will now test this fact empirically. Transform

the ramp spike time observations using the logarithm and then obtain the MLE of the parameters for a Gaussian distribution using the transformed data. Check that the results you obtain are equal to the MLE estimates obtained numerically in point 4.1.

Ex 4.3 Find now the MLE estimates for the parameters of the log-normal distribution using only the male human observations and the female human observations. Plot the two obtained log-normal densities in the same plot.

4 Molecular evolution, Jukes-Cantor model

Let the random variable X denote the nucleotide at a given position in the genome. Let the random variable Y denote the nucleotide at the same position after a certain time t (e.g. one year). Both X and Y are (extended) random variable that take values in the set $\{A, T, C, G\}$. The probability distribution of Y given X describes the mutation rate in the molecular evolution.

The Jukes-Cantor model gives the mutation probability (as a function of t) in an exponential decay manner, where:

$$P(Y = y|X = x) = \begin{cases} 0.25 + 0.75\exp(-4\alpha t), & \text{if } x = y \\ 0.25 - 0.25\exp(-4\alpha t), & \text{if } x \neq y \end{cases}$$

with $\alpha \geq 0$ being a parameter and $t > 0$ being a known constant time.

The joint probability of observing $(X = x, Y = y)$ is

$$P(X = x, Y = y) = P(Y = y|X = x)P(X = x)$$

And we suppose all nucleotides have equal marginal probabilities, such that

$$P(X = x) = 0.25 \quad x \in \{A, T, C, G\}$$

Ex 5

Ex 5.1 Investigate the behaviour of the Jukes-Cantor model for different values of α, and for $t \to 0$ and $t \to \infty$. You can implement the conditional probability of Y given X in an R function and plot it (as a function of t) for different values of α

Ex 5.2 Suppose we observe n pairs ofi i.i.d. nucleotides

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

The pairs are observed all at the same time difference t (a known constant). Write the log-likelihood function for the observations as a function of α. It can be helpful to reparameterize the observations. In the Jukes-Cantor model, we are only interested if $x = y$ or not. So we can introduce two *statistics* (functions of the observations):

$$n_1 = |\{i : X_i = Y_i\}| \quad n_2 = |\{i : X_i \neq Y_i\}|$$

It is easy to see that we can write the likelihood with respect to n_1 and n_2.

Ex 5.3 Try to find the theoretical maximum likelihood estimator for α

Ex 5.4 (if you have time, or unable to solve 5.3) Try to implement the Jukes-Cantor model in R:

- Implement the probability function
- Implement a sampling procedure to generate data accordingly
- Write the minus log-likelihood function
- Try to solve the MLE numerically