Exercises for STK4130.

Exercise 5 In a genetic model individuals belong to one out of four classes with probabilites $p_1 = \frac{1}{2} + \frac{\theta}{4}, p_2 = \frac{1}{2}(1-\theta) = p_3, p_4 = \frac{\theta}{4}$. Here θ is an unknown parameter. Assume we observe n independent individuals where X_j individuals belong to class j.

a) Argue that (X_1, X_2, X_3, X_4) has a multinomial distribution with n trials and probabilites (p_1, p_2, p_3, p_4). Find a likelihood equation for estimation of θ and show that it can be written as a quadratic equation.

b) An alternative method for finding the MLE is using the EM-algorithm by augmenting the data to $(Y_1, Y_2, Y_3, Y_4, Y_5)$ that are multinomial distributed with n trials and probabilites $(q_1, q_2, q_3, q_4, q_5) = (\frac{1}{2}, \frac{\theta}{4}, \frac{1}{2}(1-\theta), \frac{1}{2}(1-\theta), \frac{\theta}{4})$. This correspond to (X_1, X_2, X_3, X_4) being incomplete data from $(Y_1, Y_2, Y_3, Y_4, Y_5)$ with $X_1 = Y_1 + Y_2$ and $X_j = Y_{j+1}$ for $j = 2, 3, 4$.

   Show there is an explicit maximum likelihood estimator for θ based on $(Y_1, Y_2, Y_3, Y_4, Y_5)$.

c) Find the conditional expectation $\mathrm{E}[Y_1 | Y_1 + Y_2 = x_1]$ and derive an EM-algorithm for estimation of θ.

d) In a data set of $n = 197$ animals one observed $(X_1, X_2, X_3, X_4) = (125, 18, 20, 34)$. Implement the EM-algorithm and find the MLE of θ iteratively. Also solve the quadratic equation in question a).

Exercise 6 Assume that a random variable X is discrete on a set $\{x_1, \ldots, x_k\}$ with unknown probabilities $p_j = \mathrm{P}(X = x_j)$.

a) Assume we observe n independent replicates of X and records Y_j as the number of these that are equal to x_j. Explain why $\hat{p}_j = \frac{Y_j}{n}$ is the MLE for p_j.

b) Assume that we have n independent replicates of X, but that these are interval censored to intervals $(u_i, v_i]$, i.e. $u_i < X_i \leq v_i$, where u_i and v_i are known numbers. Find the conditional probabilities $\mathrm{P}(X = j | u < X \leq v)$.

c) Derive the EM-algorithm for finding the MLE of the p_j's with interval censored data.

d) For right censored data we will either know the value of X_i exactly or have a right censored data point where it is only known that $X_i \in (u_i, \infty)$. for some known value u_i. Specialize the EM-algorithm of question d) to this situation (still under the discrete model of question a).

e) Assume now a general (continuous) distribution $F(x)$ for the X_i. For right censored data it can be shown that the NPMLE (non-parametric likelihood estimator) has positive mass only in the exact observed values of the X_i's and at infinity if the largest recorded value is right censored. Furthermore, the Kaplan-Meier estimator is NPMLE. It can be expressed as

$$\hat{F}(x) = 1 - \prod_{x_j < x} [1 - \frac{Y_j}{n_j}]$$

where x_j are the points with exact observed $X_i = x_j$, Y_j the number of exact observed $X_i = x_j$ and n_j the number of individuals we know have $X_i \geq x_j$, i.e. those with exact observed $X_i \geq x_j$ and those with lower limit of the right censoring interval u_i that is greater than x_j.

In the following (mini) dataset (from Aalen, Borgan & Gjessing, Suvival and Event History Analysis, Springer, 2008) the right-censored values are indicated by astrix (*) and the exact observed values without.

2.70, 3.50*, 3.80, 4.19, 4.42, 5.43, 6.32*, 6.46*, 7.32, 8.11*

Calculate the Kaplan-Meier estimator using the formula and also using the EM-algorithm modified from question d).

Exercise 7 Assume that X_i are i.i.d. from a one-parameter exponential family with density $f(x;\theta) = \exp(\theta x - d(\theta) + S(x))$. Assume that the incomplete data are interval censored in intervals $(u_i, v_i]$ (that is $u_i < X_i \leq v_i$). Show that the EM-algorithm is based on estimating the complete data X_i conditional expectation $\mathrm{E}[X_i | u_i < X_i \leq v_i]$.