BCB BCB/GDCB/STAT/COM S 568 Spring 2011

Homework 5

February 24, 2011

Let $X = X_1 X_2 \ldots X_N$ be a sequence of scores derived from independently identically distributed random variables X_i for which $\Pr\{X_i = s_j\} = p_j$, $j = 1, 2, \ldots r$ with the restrictions $\text{Prob}\{X_i > 0\} > 0$ and $E[X] = \sum_{k=1}^{r} p_k s_k < 0$. The maximal segmental score S in X approximately follows an extreme value distribution such that

$$\Pr\{S > \frac{\ln N}{\lambda} + x\} = 1 - e^{-Ke^{-\lambda x}},$$

where λ is the unique positive root of $\text{E}[e^{\lambda X_i}] = 1$, and K is a function of λs_i.

(i) Describe how to graphically determine S and the corresponding segment.

(ii) Determine x_c such that $\Pr\{S > \frac{\ln N}{\lambda} + x_c\} = p$.

(iii) For a scoring scheme $t_i = \rho s_i$, determine the equivalent offset x_c^* giving the same probability p, i.e. find x_c^* such that $\Pr\{T > \frac{\ln N}{\lambda'} + x_c^*\} = p$, where T is the maximal segmental score and λ' is the parameter for the t_i scoring scheme.

(iv) Explain why λ can be interpreted as a scale factor.

(v) The threshold value S_p for the maximal segmental score to be significant at the p-level is $S_p = \frac{\ln N}{\lambda} + x_c$ with x_c determined as in (ii). For $\lambda = \frac{\ln 2}{2}$, determine the p-level threshold S_p' when considering a sequence of length $N' = 2N$.

(vi) Altschul (1998; Proteins 32:88) defines a normalized score as

$$S' = \frac{\lambda S - \ln K}{\ln 2}.$$

Making use of the result that the number of separate high-scoring segments, i.e. segments with scores exceeding $\frac{\ln N}{\lambda} + x$, is closely approximated by a Poisson distribution with parameter $K \exp\{-\lambda x\}$, prove his assertion that the expected number of distinct segment pairs with normalized score greater than or equal to y is well approximated by the formula

$$E(S' \geq y) \sim \frac{N}{2^y}.$$

Solution:

(i) The maximal segmental score corresponds to the highest peak in the excursion plot of E_k versus k, where $E_0 = 0$ and $E_k = \max\{E_{k-1} + X_k, 0\}$, and the coordinates of the maximal scoring segment are from the beginning of the excursion (first positive scoring position of the excursion) to the position where the peak is achieved.

(ii) We look for the solution of $1 - e^{-Ke^{-\lambda x_c}} = p$, which after a little bit of manipulation is seen to be

$$x_c = \frac{\ln K - \ln\left[\ln \frac{1}{1-p}\right]}{\lambda}.$$

(iii) By definition, λ is the unique positive root of $\text{E}[e^{\lambda X_i}] = 1$. In the t_i scoring scheme, all the X_i are multiplied by ρ, and thus λ' is seen to be $\frac{\lambda}{\rho}$. As $T = \rho S$, it is clear that the solution is $x_c^* = \rho x_c$.

(iv) As K is a function of the λs_i and by result (iii), we can multiply the s_i scores by a factor ρ, and all we would need to change in the formulae is to replace λ by $\lambda' = \frac{\lambda}{\rho}$. Equivalently, we could select a particular λ' value and scale given scores s_i by the appropriate ρ factor.

(v) The centering value $\frac{\ln N}{\lambda}$ becomes $\frac{\ln N'}{\lambda} = \frac{2 \ln 2N}{\ln 2} = \frac{\ln N}{\lambda} + 2$. Thus, $S'_p = S_p + 2$.

(vi) Subtract $\frac{\ln K}{\lambda}$ from both sides of the inequality $S > \frac{\ln N}{\lambda} + x$ and multiply by $\frac{\lambda}{\ln 2}$ to get

$$\frac{\lambda S - \ln K}{\ln 2} > \frac{\lambda}{\ln 2}[\frac{\ln N}{\lambda} + x] - \frac{\ln K}{\ln 2},$$

or $S' > y$ where $y = \frac{\ln N - \ln K}{\ln 2} + \frac{\lambda}{\ln 2}x$. Solving for x gives $x = \frac{1}{\lambda}[y \ln 2 - \ln N + \ln K]$. Inserting into $\exp\{-\lambda x + \ln K\}$ gives $\exp\{-y \ln 2 + \ln N\} = \frac{N}{2^y}$. Thus, $Prob\{S' > y\} = 1 - \exp\{-\frac{N}{2^y}\}$, and the assertion holds by the cited Poisson approximation.