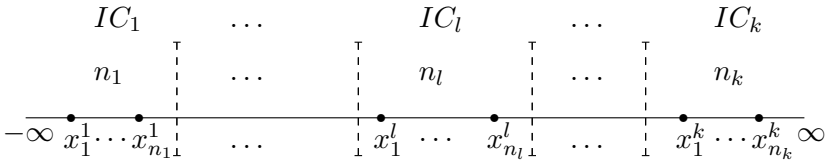
Introduction to Nonparametric Predictive Inference - Exercises

Frank Coolen & Tahani Coolen-Maturi
Department of Mathematical Sciences
Durham University, Durham, DH1 3LE, UK

SIPTA Summer School - 3 September 2010

1 Ordinal Data

Ordinal data are categorical data with a natural ordering of the categories, for example answers to questionnaires in categories such as 'very bad', 'bad', 'average', 'good', 'very good'. Suppose that there are $k \geq 2$ categories, denoted by $C_1 < C_2 < \ldots < C_k$. Let $n_l \geq 0$ be the number of observations in category C_l, for $l = 1, \ldots, k$, with $\sum_{l=1}^{k} n_l = n$. Let Y_{n+1} denote the random quantity representing the category a future observation will belong to. Using a latent variable representation, we assume that category C_l is represented by interval IC_l, with the intervals $IC_1, \ldots, IC_k$ forming a partition of the real-line and logically ordered, and the observations represented by latent x-values in the respective intervals. Thus, the following figure gives the assumed latent variable representation for the ordinal data:



The NPI lower and upper probabilities for $Y_{n+1} \in \mathcal{C}_L$, with $\mathcal{C}_L = \bigcup_{l \in L} C_l$ and $L \subset \{1, \ldots, k\}$, are derived using $A_{(n)}$ with the assumed latent variable representation, so by deriving the NPI lower and upper probabilities for the corresponding latent variable $X_{n+1} \in \mathcal{IC}_L$, where $\mathcal{IC}_L = \bigcup_{l \in L} IC_l$ and I_j are the $n+1$ intervals between the (latent) x_j^l,

$$\underline{P}(Y_{n+1} \in \mathcal{C}_L) = \underline{P}(X_{n+1} \in \mathcal{IC}_L) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}\{I_j \subset \mathcal{IC}_L\}$$

$$\overline{P}(Y_{n+1} \in \mathcal{C}_L) = \overline{P}(X_{n+1} \in \mathcal{IC}_L) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}\{I_j \cap \mathcal{IC}_L \neq \emptyset\}$$

Question A

Derive closed-form expressions for these NPI lower and upper probabilities for the special case with $\mathcal{C}_L$ consisting of neighbouring categories. So, let $\mathcal{C}_{s,t} = \bigcup_{l=s}^{t} C_l$, $\mathcal{IC}_{s,t} = \bigcup_{l=s}^{t} IC_l$ and $n_{s,t} = \sum_{l=s}^{t} n_l$, and derive $\underline{P}(Y_{n+1} \in C_{s,t})$ and $\overline{P}(Y_{n+1} \in C_{s,t})$.

Hint: Categories C_1 and C_k are different from the other categories.

Question B

Verify the expressions (1) and (2) below for $\underline{P}\,(Y_{n+1} \in \mathcal{C}_L)$ and $\overline{P}\,(Y_{n+1} \in \mathcal{C}_L)$ for the special case with $n_l > 0$ for all $l \in \{1, \ldots, k\}$. Here,

$$p_L = \sum_{r=1}^{k-1} \mathbf{1}\{r, r+1 \in L\}$$

is the number of neighbouring pairs of categories included in $\mathcal{C}_L$, and

$$e_L = \mathbf{1}\{1 \in L\} + \mathbf{1}\{k \in L\} + p_L$$

We further introduce the notation s_L for the number of categories in $\mathcal{C}_L$, so $s_L = |L|$, and $n_L = \sum_{l \in L} n_l$. The NPI lower probability, with L a strict subset of $\{1, \ldots, k\}$, is

$$\underline{P}\,(Y_{n+1} \in \mathcal{C}_L) = \underline{P}\,(X_{n+1} \in \mathcal{IC}_L) = \frac{\sum_{l \in L}(n_l - 1) + e_L}{n+1} = \frac{n_L - s_L + e_L}{n+1} \tag{1}$$

and the corresponding NPI upper probability is

$$\overline{P}\,(Y_{n+1} \in \mathcal{C}_L) = \overline{P}\,(X_{n+1} \in \mathcal{IC}_L) = \frac{\sum_{l \in L}(n_l + 1) - p_L}{n+1} = \frac{n_L + s_L - p_L}{n+1} \tag{2}$$

Question C

Suppose there are $k = 5$ ordered categories, $C_1 < \ldots < C_5$, and $n = 11$ observations with $n_1 = 1$, $n_2 = 3$, $n_3 = 1$, $n_4 = 4$ and $n_5 = 2$. Calculate the NPI lower and upper probabilities for $Y_{12} \in \mathcal{C}_L$ for the index sets L in the following table (where also n_L, the number of observations in the corresponding categories is given)

L	n_L
$\{1\}$	1
$\{2\}$	3
$\{3\}$	1
$\{4\}$	4
$\{5\}$	2
$\{1,2\}$	4
$\{1,2,3\}$	5
$\{2,3,4\}$	8
$\{1,2,4\}$	8
$\{1,2,4,5\}$	10

2 Order Statistics of Future Observations

Suppose that we have real-valued ordered data $x_1 < x_2 < \ldots < x_n$, and let $I_j = (x_{j-1}, x_j)$ for $j = 1, \ldots, n+1$ with $x_0 = -\infty$ and $x_{n+1} = \infty$. We are interested in $m \geq 1$ future observations, X_{n+i} for $i = 1, \ldots, m$. We link the data and future observations via Hill's assumption $A_{(n)}$, actually via $A_{(n+m-1)}$ which implies $A_{(n+k)}$ for all $k = 0, 1, \ldots, m-2$. Let $X_{(r)}$, for $r = 1, \ldots, m$, be the r-th ordered future observation, so $X_{(r)} = X_{n+i}$ for one $i = 1, \ldots, m$ and $X_{(1)} < X_{(2)} < \ldots < X_{(m)}$.

Question A

Show that

$$P(X_{(r)} \in I_j) = \binom{j+r-2}{j-1} \binom{n-j+1+m-r}{n-j+1} \binom{n+m}{n}^{-1}$$

Hint: Use the fact that, assuming $A_{(n)}, \ldots, A_{(n+m-1)}$ for NPI for m future observations given n real-valued observations, all orderings of these $n+m$ observations are equally likely.

Question B

Suppose we have two independent groups of real-valued observations, say X and Y, and their ordered observed values are $x_1 < x_2 < \ldots < x_{n_x}$ and $y_1 < y_2 < \ldots < y_{n_y}$, respectively. Let $I^x_{j_x} = (x_{j_x-1}, x_{j_x})$ and $I^y_{j_y} = (y_{j_y-1}, y_{j_y})$. We are interested in $m \geq 1$ future observations from each group, and want to compare the r-th ordered future observations of these groups.

Explain in detail how the following NPI lower and upper probabilities are derived:

$$\underline{P}(X_{(r)} < Y_{(r)}) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\{x_{j_x} < y_{j_y-1}\} P(X_{(r)} \in I^x_{j_x}) P(Y_{(r)} \in I^y_{j_y})$$

$$\overline{P}(X_{(r)} < Y_{(r)}) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\{x_{j_x-1} < y_{j_y}\} P(X_{(r)} \in I^x_{j_x}) P(Y_{(r)} \in I^y_{j_y})$$

Calculate these NPI lower and upper probabilities, with $m = 5$ and for all values $r = 1, 2, 3, 4, 5$, for the following two cases and comment on your results:

1. X: 1,4; Y: 2,3

2. X: 1,2,7,8; Y: 3,4,5,6

3 System Reliability

k-out-of-m systems, also called 'voting systems', are systems consisting of m components of a single type, with all components exchangeable with regard to failure or success when they need to function. The system functions if at least k of its components function. Suppose that a further n such components have been tested (so these are not the components in the system, but exchangeable with those), of which s functioned successfully.

Question A

Show that, for $0 < s < n$, the NPI upper and lower probabilities for system functioning are

$$\begin{aligned}
\overline{P}(m : k|n, s) &= \binom{n+m}{n}^{-1} \left[\binom{s+k}{s} \binom{n-s+m-k}{n-s} + \sum_{l=k+1}^{m} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right] \\
\underline{P}(m : k|n, s) &= 1 - \binom{n+m}{n}^{-1} \left[\sum_{l=0}^{k-1} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right]
\end{aligned}$$

Also derive the NPI upper and lower probabilities for system functioning for the cases $s = n$ and $s = 0$, so with all tested components functioning or all not functioning, respectively. Briefly comment on your results.

Question B

Show that for a series system, in which all components must function for the system to function (so $k = m$),

$$\begin{aligned}
(\underline{P}, \overline{P})(m : m|n, 0) &= \left(0, \prod_{j=1}^{m} \frac{j}{n+j} \right) \\
(\underline{P}, \overline{P})(m : m|n, s) &= \left(\prod_{j=1}^{m} \frac{s-1+j}{n+j}, \prod_{j=1}^{m} \frac{s+j}{n+j} \right) \quad \text{for } 0 < s < n \\
(\underline{P}, \overline{P})(m : m|n, n) &= \left(\frac{n}{n+m}, 1 \right)
\end{aligned}$$

Question C

Consider a 10-out-of-10 (series) system. Derive the NPI lower and upper probabilities that this system functions following testing of $n = 2$ components, for all possible test results ($s = 0, 1, 2$). Comment on these values, in particular discuss whether or not they are in line with your intuition (e.g. is the system more or less reliable than you had expected?).

(Challenge!) Give the NPI lower and upper probabilities for a 1-out-of-10 (parallel) system for the test results used in this example, using the results for the series system without further detailed calculation. Also comment on these values.