Nonparametric predictive inference for subcategory data

R.M. Baker, P. Coolen-Schrijner, F.P.A. Coolen
Durham University
Durham, UK
r.m.baker@dunelm.org.uk; frank.coolen@durham.ac.uk

T. Augustin
Ludwig-Maximilians University
Munich, Germany
thomas@stat.uni-muenchen.de

Abstract

Nonparametric predictive inference (NPI) is a framework for statistical inference in the absence of prior knowledge. We present NPI for multinomial data with subcategories, motivated by the hierarchical structure of many multinomial data sets. We consider situations with known and with unknown numbers of subcategories, and present lower and upper probabilities for general events involving one future observation. We present properties of the model and an algorithm to derive an approximation to the maximum entropy distribution.

Keywords. classification, multinomial data, nonparametric predictive inference, subcategories

1 Introduction

Nonparametric predictive inference (NPI) was presented by Coolen and Augustin [5, 7] for multinomial data in the absence of prior knowledge. A key assumption underlying the model is that the different categories are not ordered or otherwise related. The model is, therefore, not suited to multinomial data sets with a hierarchical structure in which two or more distinct categories may also be considered as subcategories of a single main category. Following the suggestion in [6], we present an extension of the NPI model for multinomial data suitable for data sets with subcategories, which we refer to as the Sub-MNPI model. As in the original NPI model for multinomial data [5, 7], we assume that there is no ordering of the main categories, and we also assume that for a single main category there is no ordering of its subcategories. Throughout the paper, categories are denoted by c_j and subcategories are denoted by s_{j,i_j}, where $s_{j,i_j} \subseteq c_j$. We assume that there are K main categories in total, and that k main categories have already been observed and are labelled $c_1, ..., c_k$. Similarly, we assume that there is a total of K_j subcategories in main category c_j, of which k_j have already been observed. Note that K and K_j may be known or unknown: these two situations are considered separately. Let n denote the total number of observations Y_1,...,Y_n in the data set, where n_j is the number of observations in main category c_j and n_{j,i_j} is the number of observations in subcategory s_{j,i_j}. Some main categories may not contain any subcategories, or may only be described at main category level, in which case we continue to denote these simply by c_j. Such categories are referred to as main-only categories, distinct from main categories which may or may not have specified subcategories.

In section 2 of this paper, we explain the probability wheel representation of the data on which the NPI model for subcategory data is based. In the following two sections, we then define the general events of interest for inference about a future observation and we present the NPI lower and upper probabilities for these events. The situation where K and K_j are known is considered in Section 3, and the situation where K and K_j are unknown is considered in Section 4. Some important properties of the model are then described in Section 5. In Section 6 we consider the application of the model to classification, and finally Section 7 provides some concluding remarks.

2 The Sub-MNPI model

The NPI approach for multinomial data is based on a variation of Hill's $A_{(n)}$ assumption [8] called circular-$A_{(n)}$ [5, 6, 7], which is an assumption of post-data exchangeability. The model uses a probability wheel representation of the data [5, 6, 7], where each of the n observations is represented by a radial line such that the wheel is partitioned into n equally-sized slices. From the circular-$A_{(n)}$ assumption we conclude that the next observation has probability $\frac{1}{n}$ of being in any given slice. The inferences made about a future observation therefore

depend upon which main category or subcategory each slice of the wheel represents, and this is determined by the key assumption that each main category and each subcategory is only allowed to be represented by one segment of the wheel, where a segment is defined as a single part of the wheel (note that the wheel is always divided radially) consisting of any number of full or partial slices. The assumption implies the following constraints:

- Two or more lines representing the same (sub)category must always be positioned next to each other on the wheel.
- Lines representing different subcategories within the same main category are always grouped together in one single segment of the wheel.
- If a slice is bordered by two lines representing the same (sub)category, it must be assigned to this (sub)category.
- A slice that is bordered by two lines representing observations in (sub)categories x and y where $x \neq y$, defined as a separating slice, may be assigned to x or to y or to an unobserved (sub)category not yet allocated to any other slice.
- Separating slices may be divided radially between multiple (sub)categories.

All possible configurations of the probability wheel are considered, and lower and upper probabilities for an event of interest are derived by respectively minimising and maximising the number of slices assigned to the event.

3 Known number of (sub)categories

When K and K_j, $j = 1, ..., K$, are known, the event of interest can be expressed generally as

$$E = \{Y_{n+1} \in \bigcup_{j \in J} c_j \cup \bigcup_{j \in J^*} \bigcup_{i_j \in I_j} s_{j,i_j}\} \qquad (1)$$

where $J \cap J^* = \emptyset$, $J \subseteq \{1, ..., K\}$, $J^* \subseteq \{1, ..., K\}$ and $I_j \subseteq \{1, ..., K_j\}$ for $j = 1, ..., K$. It should be emphasized that J is the index-set of the categories which occur in the event of interest only at main category level, while J^* is the index-set of the categories which occur in this event at subcategory level. We also define $\overline{I_j} = \{1, ..., K_j\} \backslash I_j$. This notation allows us to describe events which contain only specific subcategories of particular main categories, whilst also retaining the possibility of considering some main categories as a whole.
We define $OJ = J \cap \{1, ..., k\}$, which is the index-set of observed main-only categories in E, and $|OJ| = r_{main}$. We also define $UJ = J \cap \{k+1, ..., K\}$, which is the index-set of unobserved main-only categories in E, and $|UJ| = l_{main}$. Similarly, $OJ^* = J^* \cap \{1, ..., k\}$, where $|OJ^*| = r_{sub}$. OJ^* is the index-set of observed main categories in E which are described at subcategory level. We also define $UJ^* = J^* \cap \{k+1, ..., K\}$, where $|UJ^*| = l_{sub}$. UJ^* is the index-set of unobserved main categories in E which are described at subcategory level. Let $r = r_{main} + r_{sub}$, and let $l = l_{main} + l_{sub}$.
Let $OI_j = I_j \cap \{1, ..., k_j\}$, where $|OI_j| = r_j$, for $j = 1, ..., K$. OI_j is the index-set of observed subcategories in E. Also let $UI_j = I_j \cap \{k_j+1, ..., K_j\}$, where $|UI_j| = l_j$, for $j = 1, ..., K$. UI_j is the index-set of unobserved subcategories in E. Let $\overline{OI_j} = \overline{I_j} \cap \{1, ..., k_j\}$, where $\left|\overline{OI_j}\right| = \overline{r_j}$, and let $\overline{UI_j} = \overline{I_j} \cap \{k_j + 1, ..., K_j\}$, where $\left|\overline{UI_j}\right| = \overline{l_j}$.
We present the NPI lower and upper probabilities for E (1). A detailed derivation of these formulae is given in [4].

3.1 Lower probability

The NPI lower probability is found by constructing a configuration of the probability wheel which minimises the number of slices assigned to E. In order to construct such a configuration, we consider how many separating slices we can assign to main categories or subcategories not in E. First, separating slices on the wheel between different observed main categories in E can be assigned to main categories that are not in E. There are $(K-r-l)$ such categories. Furthermore, if we have subcategories which are not in E but which are part of a main category that appears in E, it may be possible to utilise these subcategories to separate observed main categories in E. By considering the configuration of the slices, we find that the number of separating slices which can potentially be filled in this way (with x^+ representing $\max\{x, 0\}$) is

$$S_M = \sum_{j \in OJ^*} \min\{(\overline{r_j} + \overline{l_j} - r_j + 1)^+, 2\} + \sum_{j \in UJ^*} \min\{\overline{l_j}, 1\}.$$

Minimising the number of slices that must be assigned to E results in the following general formula:

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \sum_{j \in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{1}{n}(2r + l - K - S_M)^+ + \frac{1}{n} \sum_{j \in OJ^*} (2r_j + l_j - K_j - 1)^+. \qquad (2)$$

Example 1 *Consider a multinomial data set with possible main categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). These main categories are labelled 1 to 6 respectively. Observations in B are further classified as light blue (LB), medium blue (MB), dark blue (DB) or other blue (OB), and observations in G are further classified as light green (LG), dark green (DG) or other green (OG). The data set consists of eight observations altogether, including 1 LB, 1 MB, 2 DB, 1 LG, 1 DG, 1 R and 1 Y.*
Suppose that we are interested in the event $Y_9 \in \{LB,MB,DB,LG,R,Y,P\}$. *We have* $K = 6$, $r = 4$ *and* $l = 1$. *For main categories described at subcategory level, the values of* K_j, r_j *and* l_j *are shown in Table 1. Here, we are unable to assign all*

	j	K_j	r_j	l_j
B	1	4	3	0
G	2	3	1	0

Table 1: Values of K_j, r_j and l_j for Example 1

separating slices within the B segment to subcategories not in E. Furthermore, we are unable to configure the probability wheel such that all observed main categories in E are separated by main categories not in E. We find that $2r+l-K = 3$ *in this example, and* $S_M = 2$. *While we can use some subcategories which are not in E but which are part of a main category that appears in E, there is still one separating slice between main categories which has to be assigned to E. Figure*

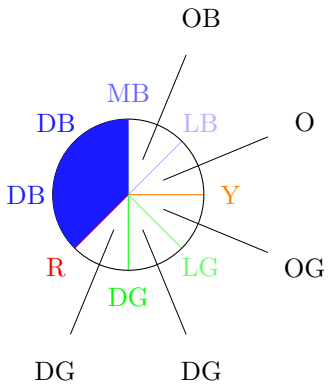

Figure 1: Probability wheel for Example 1

1 shows a possible configuration of the wheel such that O separates B and Y, OG separates Y and G, and DG separates G and R. There is then no way of separating R and B by a main category or subcategory not in E, and we are therefore forced to assign this slice to E. Looking specifically at the B segment, we see that OB separates LB and MB but the slice between MB and DB then has to be assigned to E. This leads to a NPI lower probability of $\frac{3}{8}$ *for the event E. This lower probability can be verified using (2). We see that the set OJ contains R and Y, the set* OJ^* *contains B and G, the set* OI_1 *contains LB, MB and DB and the set* OI_2 *contains LG. Also,* $2r + l - K = 3$, $S_M = 2$ *and* $\sum_{j\in OJ^*} \max\{2r_j+l_j-K_j-1, 0\} = 1$, *therefore (2) gives* $\underline{P}(E) = \frac{3}{8}$.

3.2 Upper probability

The NPI upper probability is found by constructing a configuration of the probability wheel which maximises the number of slices assigned to E. We do this by considering which slices can definitely not be assigned to E and are accounted for by the $k - r$ observed main categories not in E or by the $\overline{r_j}$ observed subcategories not in E. In order to construct such a configuration, we consider the various ways in which we can separate lines or segments on the wheel representing different main categories which either are not in E or which are present in E but have neither end of their segment in E.
First, we could separate these main categories using unobserved main categories in E. There are l of these categories. Secondly, we could separate using observed main-only categories in E. There are r_{main} such categories. Finally, we could separate using the other observed main categories in E, provided that the configuration of the relevant segment is such that each end represents a subcategory in E. There are r_{sub} main categories in E that are described at subcategory level. For a segment to have the required configuration, the category must satisfy $k_j - r_j + 1 \leq r_j + l_j$. This is because we need $k_j - r_j - 1$ subcategories in E to ensure that all subcategories not in E are separated, and a further two to ensure that both ends of the segment are in E. We define the number of main categories which satisfy this condition as $\tilde{r}_{sub}$. We define the number of main categories which are present in E but have neither end of their segment belonging to E, i.e. the number which satisfy $k_j - r_j - 1 \geq r_j + l_j$, as r^0_{sub}.
By maximising the number of slices that may be assigned to E, we find that

$$\overline{P}(E) = \sum_{j\in OJ} \frac{n_j - 1}{n} + \sum_{j\in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{\min\{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}, k\}}{n} + \sum_{j\in OJ^*} \frac{\min\{2r_j + l_j, k_j - 1\}}{n}. \quad (3)$$

Example 2 *Consider the data set described in Example 1. Suppose that we are interested in the event*

$Y_9 \in \{LB,DB,P\}$. *We have* $k = 4$, $r_{main} = 0$, $r_{sub} = 1$, $r = 1$ *and* $l = 1$. *For main categories described at subcategory level, the values of* k_j, r_j *and* l_j *are shown in Table 2. Here, we find that* $(k - r) + r^0_{sub} > l + r_{main} + \tilde{r}_{sub}$, *i.e. there is no configuration of the probability wheel such that all of the categories not in* E *are separated by a category in* E. *Also, within the* G *segment we cannot assign all separating slices to subcategories in* E. *One configuration of the wheel*

	j	k_j	r_j	l_j
B	1	3	2	0
G	2	2	0	0

Table 2: Values of k_j, r_j and l_j for Example 2

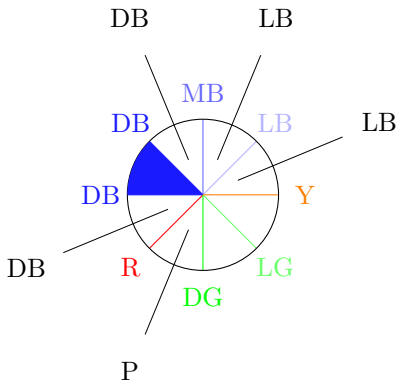

Figure 2: Probability wheel for Example 2

corresponding to the NPI upper probability is shown in Figure 2. Figure 2 shows a configuration where R *and* Y *are separated by* B, *and* G *and* R *are separated by* P. *However, we cannot separate* G *and* Y *by a category in* E. *We also do not have an available subcategory in* E *to which we can assign the slice separating* DG *and* LG. *This leads to a NPI upper probability of* $\frac{6}{8}$ *for the event* E. *This upper probability can be verified using (3).We see that the set* OJ *is empty, the set* OJ^* *contains* B *and the set* OI_1 *contains* LB *and* DB. *Also,* $r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub} = 3$ *and* $\sum_{j \in OJ^*} \min\{2r_j + l_j, k_j - 1\} = 2$, *therefore (3) gives* $\overline{P}(E) = \frac{6}{8}$.

4 Unknown number of (sub)categories

In addition to K and K_j being unknown, it is important to note that they are not assumed to have a finite limit. In order to describe the general events of interest in this situation, we introduce some new notation. Let c_{j_s}, $s = 1, ..., r'$, be the observed main-only categories in the event of interest, let UN be the set of Unobserved New main categories, which refers to any not yet observed category, and let DN_j, $j = 1, ..., l$, be the set of Defined New main categories, which is a subset of UN and which represents categories we wish to specify in the event of interest but have not yet observed.
Also, let c_{j_s}, $s = r' + 1, ..., r$, be the observed main categories in the event of interest which are described at subcategory level, and let $s_{j_s,i_{j_s}}$, $s = r' + 1, ..., r$, $i_{j_s} = 1, ..., r_s$, be the observed subcategories in the event of interest. Let $\tilde{DN}_{j_s,i_{j_s}}$, $i_{j_s} = 1, ..., d_s$, be the set of Defined New subcategories within the observed main categories c_{j_s}, and let DN_{j,i_j}, $j = 1, ..., l$, $i_j = 1, ..., l_j$, be the set of Defined New subcategories within the Defined New main categories. Let $\tilde{UN}_{j_s}$, $s = 1, ..., r$ be the set of all Unobserved New subcategories within the observed main categories c_{j_s}, and let UN_j, $j = 1, ..., l$ be the set of all Unobserved New subcategories within the Defined New main categories. A given event can be expressed as a union involving some or all of the above terms. Let $A, B \subseteq \{1, ..., k\}$ such that $A \cap B = \emptyset$. Any event of interest can be expressed using one of the two formulae shown below. The first general event is

$$
\begin{aligned}
Y_{n+1} \in & \bigcup_{s=1}^{r'} c_{j_s} \cup \bigcup_{s=r'+1}^{r} \Big(\bigcup_{i_{j_s}=1}^{r_s} s_{j_s,i_{j_s}} \Big) \\
& \cup \bigcup_{s \in A} \Big(\tilde{UN}_{j_s} \setminus \bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\
& \cup \bigcup_{s \in B} \Big(\bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\
& \cup \bigcup_{j=1}^{l'} \Big(UN_j \setminus \bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \cup \bigcup_{j=l'+1}^{l} \Big(\bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big).
\end{aligned} \tag{4}
$$

The second general event is

$$
\begin{aligned}
Y_{n+1} \in & \bigcup_{s=1}^{r'} c_{j_s} \cup \bigcup_{s=r'+1}^{r} \Big(\bigcup_{i_{j_s}=1}^{r_s} s_{j_s,i_{j_s}} \Big) \\
& \cup \bigcup_{s \in A} \Big(\tilde{UN}_{j_s} \setminus \bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \cup \bigcup_{s \in B} \Big(\bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\
& \cup UN \setminus \Big\{ \bigcup_{j=1}^{l'} \Big(UN_j \setminus \bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \\
& \cup \bigcup_{j=l'+1}^{l} \Big(\bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \Big\}.
\end{aligned} \tag{5}
$$

We denote these by E_1 (4) and E_2 (5). We now present formulae for the NPI lower and upper

probabilities for each of these general events. A detailed derivation of these formulae is given in [4].

4.1 Lower probability

First we consider event E_1, which includes only a finite number of unobserved main categories. By minimising the number of slices of the wheel that must be assigned to E_1, the NPI lower probability is

$$\underline{P}(E_1) = \sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n})\} + \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n}) + \frac{N_s}{n} \} \tag{6}$$

where $N_s = [(r_s - 1) - d_s - (k_{j_s} - r_s)]^+$.
For E_2, which contains all except a finite number of the UN main categories, the NPI lower probability is

$$\underline{P}(E_2) = \sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n})\} + \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n}) + \frac{N_s}{n} \} + \frac{1}{n}(2r - k - l - \sum_{s=r'+1}^{r} M_s)^+ \tag{7}$$

where $M_s = 2$ if $s \notin A$ and $M_s = \min\{[d_s + (k_{j_s} - r_s) - (r_s - 1)]^+, 2\}$ if $s \in A$.

Example 3 *Consider a multinomial data set where the set of possible main categories consists of an unknown number of different colours. We have observed the following main categories: red (R), blue (B), green (G) and pink (P). At subcategory level, we have observed DB, MB, LB, DG, MG, LG, MP and DP. In addition we define two new main categories: orange (O), with defined subcategories LO and MO, and purple (Pu) with defined subcategory DPu. We also define the new subcategory LP. We let UN_B represent all unobserved new subcategories within the main category B, including the defined new subcategory RB, and let UN_{Pu} represent the equivalent for the main category Pu. The data set consists of twenty observations including 3 R, 3 DB, 1 MB, 2 LB, 3 DG, 2 MG, 2 LG, 2 MP and 2 DP. We consider the event $Y_{21} \in \{(LB \cup MB) \cup (LG \cup MG) \cup (MP) \cup (UN_B \backslash RB) \cup (LP) \cup [UN \backslash ((UN_{Pu} \backslash DPu) \cup (LO \cup MO))]\}$. We label this event E. Let $s = 1$ correspond to B, $s = 2$ to G and $s = 3$ to P. This is an event of type E_2 (5), so (7) is used to compute the NPI lower probability for this event.*
In this example, $r = 3$. The main categories for which $s \notin A$ are G and P, and the only main category for which $s \in A$ is B. We have $N_1 = [(r_1 - 1) - d_1 - (k_{j_1} - r_1)]^+ = [(2 - 1) - 1 - (3 - 2)]^+ = 0$. We also have $M_1 = \min\{[d_1 + (k_{j_1} - r_1) - (r_1 - 1)]^*, 2\} = 1$, $M_2 = 2$ and $M_3 = 2$. Therefore $\sum_{s=1}^{3} M_s = 5$. The values of n_{j_s} and $n_{j_s,i_{j_s}}$ are shown in Tables 3 and 4.

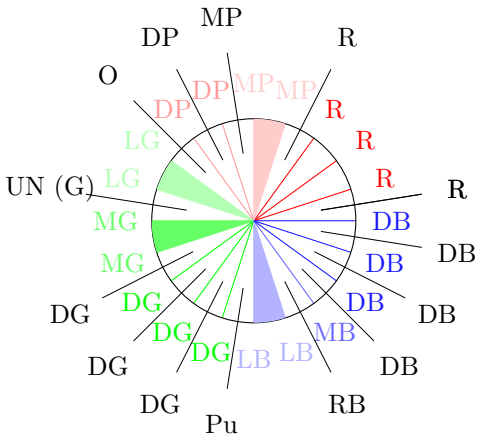

Figure 3: Probability wheel for Example 3

	B	G	P
n_{j_s}	6	7	4

Table 3: Values of n_{j_s} for Example 3

	LB	MB	LG	MG	MP
$n_{j_s,i_{j_s}}$	2	1	2	2	2

Table 4: Values of $n_{j_s,i_{j_s}}$ for Example 3

By (7), the NPI lower probability for the event E is $\frac{4}{20}$. Figure 3 shows a corresponding configuration of the probability wheel. There are four slices assigned to E, and the remaining slices are assigned to main categories or subcategories not in E and are labelled accordingly.

4.2 Upper probability

The NPI upper probabilities for events E_1 and E_2 are derived by assigning as many slices of the wheel as possible to the event of interest. The NPI upper

probability for event E_1 is

$$\overline{P}(E_1) = \sum_{s=1}^{r'} \frac{n_{j_s}-1}{n} + \sum_{s\notin A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s,i_{j_s}}-1}{n}) + \frac{k_{j_s}-1-P_s}{n}\} + \frac{\min\{r-r^0+l+\tilde{r},k\}}{n} + \sum_{s\in A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s,i_{j_s}}-1}{n}) + \frac{k_{j_s}-1}{n}\} \quad (8)$$

where $P_s = [(k_{j_s} - r_s - 1) - r_s - d_s]^+$, r^0 denotes the number of main categories such that $s \notin A$ which satisfy $r_s + d_s - (k_{j_s} - r_s - 1) \leq 0$, and $\tilde{r}$ denotes the number of main categories c_{j_s} which satisfy either $s \in \{1, ..., r'\}$, $s \in A$ or the condition

$$s \notin A, \qquad r_s + d_s - (k_{j_s} - r_s - 1) \geq 2.$$

The NPI upper probability for event E_2 is

$$\overline{P}(E_2) = \sum_{s=1}^{r'} \frac{n_{j_s}-1}{n} + \sum_{s\notin A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s i_{j_s}}-1}{n}) + \frac{k_{j_s}-1-P_s}{n}\} + \frac{k}{n} + \sum_{s\in A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s i_{j_s}}-1}{n}) + \frac{k_{j_s}-1}{n}\}. \quad (9)$$

Example 4 *Consider the data set described in Example 3. Suppose that we are interested in the event* $Y_{21} \in \{(LB \cup MB) \cup (LG \cup MG) \cup (MP) \cup (UN_B \backslash RB) \cup (LP) \cup (UN_{Pu} \backslash DPu) \cup (LO \cup MO)\}$. *We label this event* E. *This is an event of type* E_1, *so (8) is used for the NPI upper probability for* E.
In this example, $r = 3$, $l = 2$ *and* $k = 4$. *Let* $s = 1$ *correspond to B,* $s = 2$ *to G and* $s = 3$ *to P. The main categories for which* $s \notin A$ *are G and P, and the only main category for which* $s \in A$ *is B. We have* $P_2 = [(k_{j_2} - r_2 - 1) - r_2 - d_2]^+ = [(3-2-1)-2-1]^+ = 0$ *and* $P_3 = [(k_{j_3} - r_3 - 1) - r_3 - d_3]^+ = [(2-2-1)-1-1]^+ = 0$. *The values of* n_{j_s}, k_{j_s} *and* $n_{j_s,i_{j_s}}$ *are shown in Tables 5 and 6.*

	B	G	P
n_{j_s}	6	7	4
k_{j_s}	3	3	2

Table 5: Values of n_{j_s} and k_{j_s} for Example 4

We have $r^0 = 0$ *and* $\tilde{r} = 3$, *as both of the main categories in E for which* $s \notin A$ *satisfy the condition* $r_s + d_s - (k_{j_s} - r_s - 1) \geq 2$. *The general formula (8) shows that the NPI upper probability for the event E is* $\frac{13}{20}$. *Figure 4 shows a corresponding configuration of the probability wheel. There are four slices of the wheel that must be assigned to E. The nine further slices that can be assigned to elements of E are labelled accordingly.*

	LB	MB	LG	MG	MP
$n_{j_s,i_{j_s}}$	2	1	2	2	2

Table 6: Values of $n_{j_s,i_{j_s}}$ for Example 4

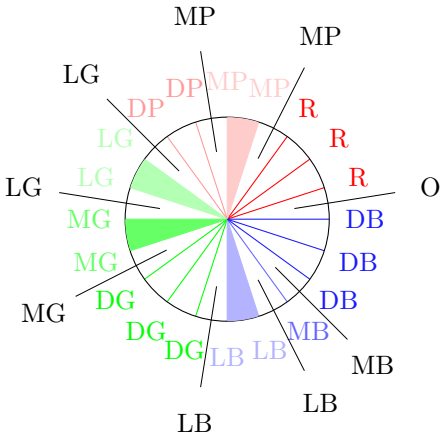

Figure 4: Probability wheel for Example 4

5 Properties of the model

We now discuss several properties of the Sub-MNPI model. We focus here on the case where K and K_j are known, but the following properties are equally applicable when these quantities are unknown.
A fundamental property for lower and upper probabilities is the conjugacy property, which states that $\overline{P}(E) = 1 - \underline{P}(E^c)$. This is implicit in the F-probability property, proven below, but can also be proven explicitly for the Sub-MNPI model [4]. It can also be shown [4] that the interval between the lower and upper probabilities always contains the relative frequency of observations in the event of interest E, i.e.

$$\underline{P}(E) \leq \sum_{j\in OJ} \frac{n_j}{n} + \sum_{j\in OJ^*}\sum_{i_j\in OI_j} \frac{n_{j,i_j}}{n} \leq \overline{P}(E). \quad (10)$$

This is an attractive property, since it shows that the Sub-MNPI model is not in conflict with the empirical probability, and one which is not always satisfied by methods such as Bayesian inferences which typically assign a positive probability to a category before it has been observed even once. A third property that can be proven [4] is that as the number of observations in the data set becomes infinitely large, the imprecision vanishes and the interval probability $P(E)$ shrinks to

a point value equal to the relative frequency. This is, in our situation, a desirable property for the model.
We now prove that the interval probabilities $[\underline{P}(E), \overline{P}(E)]$ given by the Sub-MNPI model are F-probabilities in the sense of Weichselberger [13]. F-probability is a desirable property, because it shows that none of the interval probabilities are too wide and that they could not be made any smaller given the data available to us. Also, F-probability is strongly linked to other concepts in imprecise probability theory. As stated above, conjugacy is implicit in the F-probability property. Coherence is a direct consequence of F-probability, by Walley's lower envelope theorem [11], and this can be seen as a rationality requirement. The following is based on work by Coolen and Augustin [7] that proved the F-probability property for the original NPI model for multinomial data.
For the proof we introduce some new notation in order to describe all the possible configurations of the probability wheel. Suppose that the wheel is split into K segments, and each segment is split into K_j subsegments. We move clockwise around the wheel numbering the segments as $1, ..., K$ as shown in Figure 5. We also number the subsegments within segment j as $1, ..., K_j$ as shown in Figure 6. The area of these

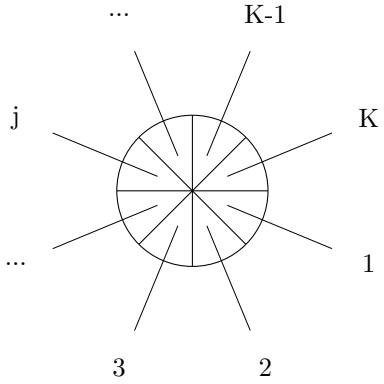

Figure 5: Numbering of segments

segments and subsegments is thus far unspecified: we allocate a different main category or subcategory to each segment or subsegment in order to describe the configuration of the wheel, but a segment assigned to an unobserved category may have area zero.
As seen in [7], we let Σ represent the set of all possible configurations σ of the wheel. Each σ can be described by a sequence

$$(\sigma(j))_{j=1...K+1}, \quad \sigma(K+1) = \sigma(1)$$

where $\sigma(j)$ is the index of the main category assigned to segment j, and a set of sequences

$$(\sigma(i,j))_{i=1...K_j}, \quad j \in J^*$$

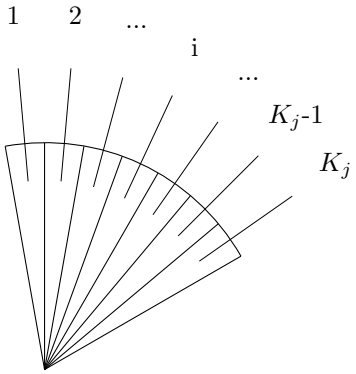

Figure 6: Numbering of subsegments

where $\sigma(i,j)$ is the index of the subcategory within main category j assigned to subsegment i.
It is also necessary to describe the position of the observed main categories and subcategories on the wheel for a given σ. Let the circular sequence

$$\sigma(i_1), ..., \sigma(i_{k+1}), \sigma(i_{k+1}) = \sigma(i_1)$$

be the indices of the observed main categories as we move around the wheel, and let the sequence

$$\sigma(i_1, j), ..., \sigma(i_{k_j}, j), \quad j \in J^*$$

be the indices of the observed subcategories as we move through the segment representing main category j.
For $l = 1, ..., k$, we describe each separating slice between two main categories as follows:

$$J_{\sigma,l} = \{\sigma(j) | i_l \leq j \leq i_{l+1}\}$$

if categories in positions i_l and i_{l+1} are main-only,

$$J_{\sigma,l} = \{\sigma(j) | i_l \leq j < i_{l+1}\} \cup \bigcup_{x=1}^{i_1} \sigma(x, l+1)$$

if category in position i_l is main-only but category in position i_{l+1} has subcategories,

$$J_{\sigma,l} = \{\sigma(j) | i_l < j \leq i_{l+1}\} \cup \bigcup_{x=i_{k_l}}^{K_l} \sigma(x, l)$$

if category in position i_l has subcategories but category in position i_{l+1} is main-only, and

$$J_{\sigma,l} = \{\sigma(j) | i_l < j < i_{l+1}\} \cup \bigcup_{x=1}^{i_1} \sigma(x, l+1) \cup \bigcup_{x=i_{k_l}}^{K_l} \sigma(x, l)$$

if categories in positions i_l and i_{l+1} both have subcategories. $J_{\sigma,l}$ is the index set of all main categories and subcategories to which the separating

slice could be assigned. Let $c_{|J_{\sigma,l}|}$ be the set of all these main categories and subcategories.
We also describe the separating slice between two observed subcategories within the same main category using

$$B_{\sigma,j,l} = \{\sigma(b,j) | i_l \leq b \leq i_{l+1}\}, l = 1, ..., k_j - 1, j \in J^*.$$

This is the set of indices of all possible subcategories to which, for the particular configuration σ, we could assign the separating slice between the subcategories in positions i_l and i_{l+1} in the segment representing main category j. Let $s_{|B_{\sigma,j,l}|}$ be the set of these subcategories.
Now, for a given configuration σ, the Sub-MNPI model gives the following basic probability assignment [3] to the event $Y_{n+1} \in c_j$:

$$m_\sigma(Y_{n+1} \in c_j) = \max\{\frac{n_j - 1}{n}, 0\}, j = 1, ..., K.$$

Similarly, the basic probability assignment given to the event $Y_{n+1} \in s_{j,i_j}$ is

$$m_\sigma(Y_{n+1} \in s_{j,i_j}) = \max\{\frac{n_{j,i_j} - 1}{n}, 0\}, i = 1, ..., K_j.$$

With regard to distributing probability mass amongst slices separating different main categories or subcategories, we give the following basic probability assignments:

$$m_\sigma(Y_{n+1} \in c_{|J_{\sigma,l}|}) = \frac{1}{n}, \quad l = 1, ..., k.$$

$$m_\sigma(Y_{n+1} \in s_{|B_{\sigma,j,l}|}) = \frac{1}{n}, \quad l = 1, ..., k_j - 1, j \in J^*.$$

Any other event is given the basic probability assignment of zero.
Let X_E represent the index set of the event of interest E. This set contains some one-dimensional elements, corresponding to main-only categories, and some two-dimensional elements, corresponding to subcategories. We now determine the lower and upper probabilities for event E via the belief and plausibility functions [10]. For a particular configuration σ, we find that the belief function of E is

$$\begin{aligned}\underline{P}_\sigma(E) = &\sum_{j \in J} m_\sigma(\{Y_{n+1} \in c_j\}) \\ &+ \sum_{j \in J^*} \sum_{i_j \in I_j} m_\sigma(\{Y_{n+1} \in s_{j,i_j}\}) \\ &+ \sum_{J_{\sigma,l} \subseteq X_E} m_\sigma(\{Y_{n+1} \in c_{|J_{\sigma,l}|}\}) \\ &+ \sum_{B_{\sigma,j,l} \subseteq I_j} m_\sigma(\{Y_{n+1} \in s_{|B_{\sigma,j,l}|}\})\end{aligned} \tag{11}$$

and the plausibility function of E is

$$\begin{aligned}\overline{P}_\sigma(E) = &\sum_{j \in J} m_\sigma(\{Y_{n+1} \in c_j\}) \\ &+ \sum_{j \in J^*} \sum_{i_j \in I_j} m_\sigma(\{Y_{n+1} \in s_{j,i_j}\}) \\ &+ \sum_{J_{\sigma,l} \cap X_E \neq \emptyset} m_\sigma(\{Y_{n+1} \in c_{|J_{\sigma,l}|}\}) \\ &+ \sum_{B_{\sigma,j,l} \cap I_j \neq \emptyset} m_\sigma(\{Y_{n+1} \in s_{|B_{\sigma,j,l}|}\}).\end{aligned} \tag{12}$$

We therefore have a set of belief functions and a set of plausibility functions corresponding to the set Σ of possible configurations of the probability wheel. According to Theorem 3.2 of [3], and to [9], taking the lower and upper envelopes over all possible configurations leads to F-probability. Since the lower and upper probability formulae of the Sub-MNPI model are derived by considering all possible configurations $\sigma \in \Sigma$, resulting in

$$\underline{P}(E) = \min_{\sigma \in \Sigma} \underline{P}_\sigma(E)$$

and

$$\overline{P}(E) = \max_{\sigma \in \Sigma} \overline{P}_\sigma(E),$$

the interval probability $[\underline{P}(E), \overline{P}(E)]$ is an F-probability.

6 Approximate maximum entropy distribution

We present an algorithm for approximating the maximum entropy distribution consistent with the Sub-MNPI model, with a view to using this maximum entropy measure in the construction of classification trees. Further details of such classification at main category level are presented in [4]; the implementation of this method at subcategory level is ongoing research.
The process of computing the maximum entropy distribution is carried out in two stages. Initially, we work at main category level only. We apply the NPI-M algorithm presented in [1], which gives a maximum entropy probability $p_{maxE}(c_j)$ for each main category. As a second step, we share the probability mass $p_{maxE}(c_j)$ as evenly as possible between the subcategories, in such a way that the probability $\widehat{p}_{j,i_j}$ that is assigned by the algorithm to subcategory s_{j,i_j} is within the interval $[L_{j,i_j}, U_{j,i_j}]$. Let $K(i)_j$ represent the number of subcategories in main category c_j that have been observed i times. From the NPI-M algorithm [1] we have the results $p_j = p_{maxE}(c_j), j = 1, ..., K$. This means that for

each main category c_j, we have a segment consisting of np_j slices. Of these slices, $n(\sum_{i=1}^{K_j} L_{j,i_j})$ must be assigned to observed subcategories in c_j. We therefore have remaining probability mass $p_j - \sum_{i=1}^{K_j} L_{j,i_j}$ that may be assigned to any available subcategory in c_j, and this is termed optional probability mass. For each c_j, we share the optional probability mass between subcategories of c_j, beginning with subcategories with the fewest observations. This leads to the Sub-A-NPI-M algorithm, which is shown below in pseudo-code and which is similar to the A-NPI-M algorithm presented in [1] and justified in the same way.

Sub-A-NPI-M

For $j = 1$ to K

For $i = 1$ to K_j

$L_{j,i_j} \leftarrow \max\{\frac{n_{j,i_j}-1}{n}, 0\}$

$opt \leftarrow p_j - \sum_{i=1}^{K_j} L_{j,i_j}$

$\widehat{p}_{j,i_j} \leftarrow L_{j,i_j}$

$t \leftarrow 0$;

While $(opt > 0)$ do

If $(n_{j,i_j} = t$ *or* $n_{j,i_j} = t+1)$ $\widehat{p}_{j,i_j} \leftarrow \widehat{p}_{j,i_j} + \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\}$;

$opt \leftarrow opt - \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\}$;

$t \leftarrow t + 1$;

The Sub-A-NPI-M algorithm is illustrated in Example 5.

Example 5 *Consider a multinomial data set with observed main categories blue (B), green (G), red (R) and pink (P), and unobserved main category orange (O). Observations in B are further classified as light blue (LB) or dark blue (DB), and observations in G are further classified as light green (LG) or dark green (DG). The data set consists of twenty observations altogether, including 5 DB, 5 DG, 5 R and 5 P.*
First, considering the data at main category level only, we apply the NPI-M algorithm [1] and find that the maximum entropy probabilities assigned to the main categories $\{O,R,B,G,P\}$ *are* $\{\frac{1}{20}, \frac{19}{80}, \frac{19}{80}, \frac{19}{80}, \frac{19}{80}\}$. *(For further details on this, see [1] and [4].) A configuration of the wheel corresponding to this distribution is shown in Figure 7. The separating slices are shared in such a way that B, R, G and P are each assigned* $\frac{3}{4}$ *of a separating slice. We now consider the subcategories. The maximum entropy probabilities for the main categories are distributed over the subcategories using the Sub-A-NPI-M algorithm. For main category B we have* $\underline{P}(DB) = \frac{4}{20}$ *and* $\underline{P}(LB) = 0$. *For main category G we have* $\underline{P}(DG) = \frac{4}{20}$ *and* $\underline{P}(LG) = 0$. *Applying the Sub-A-NPI-M algorithm, we find that* $opt = \frac{19}{80} - \frac{4}{20} = \frac{3}{80}$ *for both of these main categories. Taking* $t = 0$ *gives*

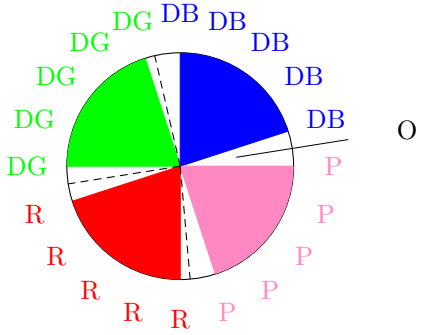

Figure 7: Probability wheel for Example 5

$$\widehat{p}(LB) = 0 + \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\} = \frac{3}{80},$$

$$\widehat{p}(LG) = 0 + \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\} = \frac{3}{80}.$$

So the probabilities assigned to the set of subcategories $\{LB,DB\}$ *are* $\{\frac{3}{80}, \frac{4}{20}\}$ *and the probabilities assigned to the set of subcategories* $\{LG,DG\}$ *are* $\{\frac{3}{80}, \frac{4}{20}\}$.

The Sub-A-NPI-M algorithm can be implemented for building classification trees using methodology similar to that shown in [2] and in [4].

7 Concluding remarks

In this paper we presented the Sub-MNPI model for inferences from multinomial data described at subcategory level as well as at main category level. NPI lower and upper probabilities were derived for the general events of interest, and some fundamental properties of the model were explained. The inferences presented here are more flexible than those given by the original NPI model for multinomial data in the sense that observations can be represented at varying levels of detail, which makes the model widely applicable to practical problems. With the view to applying the Sub-MNPI model to classification problems, an algorithm was presented for approximating the maximum entropy distribution consistent with these inferences. Implementation of this algorithm for building classification trees, and

comparison of the approach with alternative imprecise and classical methods, is ongoing. It is also of interest for future research to investigate other applications of the Sub-MNPI model.

With regard to future research, it will also be useful to compare classification trees built using the Sub-A-NPI-M algorithm presented here with classification trees constructed by ignoring the hierarchical relationship between the categories and subcategories and simply using the NPI-M algorithm presented in [1]. Note that the distinction between these two methods, and the different results they achieve, show that the Representation Invariance Principle (RIP) satisfied by Walley's IDM [12] does not generally hold for NPI. This is an issue discussed in detail by Coolen and Augustin [5, 7].

The Sub-MNPI model presented in this paper could be extended further by considering inferences about multiple future observations and by introducing further layers e.g. subsubcategories to the hierarchy. Such developments would be of theoretical and practical interest.

Acknowledgements

We thank two referees for supportive comments and suggestions to improve the presentation of this paper.

References

[1] Abellán, J., Baker, R.M. and Coolen, F.P.A. (2011) Maximising entropy on the nonparametric predictive inference model for multinomial data. *European Journal of Operational Research*, **212**, 112-122.

[2] Abellán, J. and Moral, S. (2003) Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, **18**, 1215-1225.

[3] Augustin, T. (2005) Generalized basic probability assignments. *International Journal of General Systems*, **34**, 451-463.

[4] Baker, R.M. (2010) Nonparametric predictive inference: Selection, classification and subcategory data, PhD thesis, Durham University. *http://etheses.dur.ac.uk/257*

[5] Coolen, F.P.A. and Augustin, T. (2005) Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. *Proceedings of the Fourth International Symposium on Imprecise Probability: Theories and Applications.*

[6] Coolen, F.P.A. and Augustin, T. (2007) Multinomial nonparametric predictive inference with subcategories. *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications.*

[7] Coolen, F.P.A. and Augustin, T. (2009) A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. *International Journal of Approximate Reasoning*, **50**, 217-230.

[8] Hill, B.M. (1968) Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677-691.

[9] Miranda, E., de Cooman, G. and Couso, I. (2005) Lower previsions induced by multi-valued mappings. *Journal of Statistical Planning and Inference*, **133**, 173-197.

[10] Shafer, G. (1976) *A Mathematical Theory of Evidence*. Princeton University Press.

[11] Walley, P. (1991) *Statistical Reasoning With Imprecise Probabilities.* Chapman and Hall.

[12] Walley, P. (1996) Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society* B, **58**, 3-57.

[13] Weichselberger, K. (2000) The theory of interval probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, **24**, 149-170.