Multinomial nonparametric predictive inference with sub-categories

F.P.A. Coolen
Durham University
Frank.Coolen@durham.ac.uk

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

Abstract

Nonparametric predictive inference (NPI) is a powerful tool for predictive inference under nearly complete prior ignorance. After summarizing our NPI approach for multinomial data, as presented in [8, 9], both for situations with and without known total number of possible categories, we illustrate how this approach can be generalized to deal with sub-categories, enabling consistent inferences at different levels of detail for the specification of observations. This approach deals with main categories and sub-categories in a logical manner, directly based on the powerful probability wheel representation for multinomial data that is central to our method and that ensures strong internal consistency properties. Detailed theory for such inferences, enabling for example more layers of sub-categories as might occur in tree-like data base structures, has yet to be developed, but is conceptually straightforward and in line with the illustrations for more basic inferences presented in this paper.

Keywords. CA model, imprecise Dirichlet model, nonparametric predictive inference, probability wheel representation.

1 Introduction

Statistical data in various application areas are often multinomial, i.e. the observations fall into one of several unordered categories. Recently, the current authors have developed a nonparametric predictive inferential approach for such data [8, 9]. This approach provides lower and upper probabilities for a future observation, on the basis of observed multinomial data, and it adds only few modelling assumptions to the data. The method has been presented both for situations in which one has no information about the number of possible categories [8], and for situations with at most K possible categories [9], where the additional knowledge in the latter case leads to less imprecision for some events of interest. In this paper, we will refer to the general NPI approach for multinomial data, by Coolen and Augustin [8, 9], as the 'CA model'[1]. In the earlier papers, the advantages of the CA model are discussed and illustrated in detail, and the resulting lower and upper probabilities are also compared to those based on Walley's Imprecise Dirichlet Model (IDM) [23], which has attracted considerable attention in a variety of application areas [4].

The CA model fits in the framework of 'Nonparametric Predictive Inference' (NPI) [2, 7], which is generally based on Hill's assumption $A_{(n)}$ [18]. However, for multinomial data, a variation of this assumption is required, which was introduced by Coolen and Augustin [8] and called 'circular-$A_{(n)}$', and which is very close in nature to Hill's $A_{(n)}$ as both are post-data versions of exchangeability [14]. Coolen [7] illustrated the natural use of circular-$A_{(n)}$ for circular data.

A key assumption for the CA model as presented before [8, 9], as well as for most models for multinomial data including Walley's IDM, is that the different categories are in no way related. Not only should the categories not be ordered, but there should also not be other possible links between some of the categories. For example, such methods are not fully suited for situations where two or more categories may be considered as sub-categories of a larger category, for example[2] one may be interested in situations where one distinguishes between main colours such as green, red and blue, but in addition distinguishes between light-blue and dark-blue within the latter category. An interesting property, called the Representation Invariance Principle (RIP), of Walley's IDM [23] is that this distinction has no effect on probabilities for events which do not directly involve 'blue', this property does not hold in general for the CA model [8, 9]. In this pa-

[1] CA: *Circulus Alearius* and/or *Circular-*$A_{(n)}$.

[2] The use of colours as different categories in illustrative examples might be considered inappropriate, as one could consider an existing natural ordering of colours, but it has become somewhat of a tradition in this field following Walley [23].

per, we call categories such as light-blue and dark-blue 'sub-categories' of the category blue, and we present the basic way in which the CA model can deal explicitly with such sub-categories.

In Section 2 of this paper, we present an overview of the CA model as presented before, both for a known and unknown number of categories [8, 9]. Section 3 illustrates how the CA model should be generalized in order to deal with sub-categories, which is mostly explained via an example as the main theory is under development, and Section 4 provides some concluding remarks.

2 The CA model

2.1 The basic setting

Hill [18] introduced the assumption $A_{(n)}$ as a basis for predictive inference in case of real-valued observations. Suppose we have n observations ordered as $z_1 < z_2 < \ldots < z_n$, which partition the real-line into $n+1$ intervals (z_{j-1}, z_j) for $j = 1, \ldots, n+1$, where we use notation $z_0 = -\infty$ and $z_{n+1} = \infty$. Hill's assumption $A_{(n)}$ is that a future observation, represented by a random quantity Z_{n+1}, falls into any such interval with equal probability, so we have $P(Z_{n+1} \in (z_{j-1}, z_j)) = \frac{1}{n+1}$ for $j = 1, \ldots, n+1$. This assumption implies that the rank of Z_{n+1} amongst the n observed data has equal probability to be any value in $\{1, \ldots, n+1\}$. This clearly is a post-data assumption, related to exchangeability [14], which provides direct posterior predictive probabilities [13]. Hill [18, 19] argued that $A_{(n)}$ is a reasonable basis for inference in the absence of any further process information beyond the data set, when actually predicting a future random quantity. Augustin and Coolen [2] prove generally that Nonparametric Predictive Inference (NPI) based on $A_{(n)}$ has strong consistency properties in the theory of interval probability [22, 24, 25]. Interestingly, as NPI is based on $A_{(n)}$, such inference is fully in line with 'perfectly calibrated' inference along the lines of Lawless and Fredette [20], who however restricted attention to precise probability.

In the CA model, multinomial data are represented as observations on a probability wheel, and hence as circular data. A straightforward variation of $A_{(n)}$ that is suitable for inference based on such data, and again linked to exchangeability of $n+1$ observations, is the assumption *circular*-$A_{(n)}$, denoted by $\textcircled{A}_{(n)}$ [7, 8]: Let ordered circular data $x_1 < x_2 < \ldots < x_n$ create n intervals on a circle, denoted by $I_j = (x_j, x_{j+1})$ for $j = 1, \ldots, n-1$, and $I_n = (x_n, x_1)$. The assumption $\textcircled{A}_{(n)}$ is that a future observation X_{n+1} falls into each of these n intervals with equal (classical) probability, so

$$P(X_{n+1} \in I_j) = \frac{1}{n}, \quad \text{for } j = 1, \ldots, n. \qquad (1)$$

Clearly, $\textcircled{A}_{(n)}$ is again a post-data assumption, related to the appropriate exchangeability assumption for such circular data, in exactly the same way as $A_{(n)}$ was related to exchangeability of $n+1$ values on the real-line. NPI based on $\textcircled{A}_{(n)}$ has the same consistency properties as shown in [2] for such inference based on $A_{(n)}$.

In the CA model [8, 9], $\textcircled{A}_{(n)}$ is combined with the assumed underlying representation of multinomial data as outcomes of spinning a probability wheel. Without additional assumptions about the probability mass $1/n$ per interval I_j, the predictive inferences based on the CA model are again in the form of interval probabilities [2, 22, 24, 25], where a lower probability for an event A is represented by $\underline{P}(A)$, and the corresponding upper probability by $\overline{P}(A)$. Effectively, the lower probability is the maximum lower bound for the classical probability for A that is consistent with the probabilities as assigned by $\textcircled{A}_{(n)}$ and in accordance with the probability wheel model, according to De Finetti's fundamental theorem of probability [14], and the upper probability is the minimum upper bound consistent in this way.

The predictive lower and upper probabilities presented in [8, 9], and reviewed in this section, are based on an underlying assumed model, ensuring that they not only make sense for one specific set of data, which they do being F-probability [24, 25] and due to the fact that they bound the observed relative frequencies, but they are also consistent if more observations are added to the data. We now give a brief summary of the key aspects of this model and its properties.

The CA model underlying the nonparametric predictive lower and upper probabilities presented below, is based on a probability wheel representation, with each observation category represented by a single segment of the probability wheel. The idea of such a probability wheel is as follows (see [15] for use of the same concept as a reference experiment underlying subjective probability). An arrow, fixed at the center of a circle, spins around, such that the arrow is equally likely to stop at any segment of the same size, where a segment is an area between two lines from the center of the circle to its circumference. In our model for multinomial data, we assume explicitly that each possible observation category is represented by only a single segment on the circle. Even more, we assume that there is no natural (or assumed) ordering of the observation categories, and therefore also no such ordering of the segments on the circle. Clearly, if we had perfect knowledge of the sizes of all segments on the probability wheel, we would have full knowledge of the probability distribution for future observations from this multinomial setting. The CA model can deal both with situations where the number of possible categories is unknown [8] and where it is known that there are K possible categories [9], and it only assumes a finite number of exchangeable multinomial observations, $\textcircled{A}_{(n)}$, and the probability wheel representation. As this probability wheel is only an abstract model, we have no information about the configuration of different segments on it. This is important for our nonparametric predictive inferences based on $\textcircled{A}_{(n)}$ once we consider unions of two or more categories, and adds to imprecision of our inferences, in the sense that our lower and upper probabilities are optimal bounds over all configurations of the possible segments on the probability wheel. In Section 3 we change this perspective a little, by allowing categories to be subdivided into sub-categories, in such a way that both inferences at the category and at the sub-category level can be considered. We will show how the CA model can deal with sub-categories by explicitly representing sub-categories within the corresponding category in the probability wheel representation. Each sub-category is again assumed to be represented by a single segment on the probability wheel.

When we combine the concept of a probability wheel, with each observation category represented by a single segment, with the assumption $\textcircled{A}_{(n)}$, on the basis of n observations, then we can represent this situation as if the n observations are represented by n lines, which partition the circle into n equally sized slices, representing that the next observation is equally likely to fall into each one of these slices. The assumption that each observation category is represented by only one segment on the probability wheel, implies that the lines representing observations in the same category are 'next to each other'. For example, if precisely two observations fall into one category, then our current inferences with regard to the next observation falling into this category, are based on the current representation with two lines next to each other which both represent this category, and the other lines, in case of more than 2 observations, representing different categories. Under the assumption $\textcircled{A}_{(n)}$, the probability $\frac{1}{n}$ for the line on the probability wheel corresponding to the next observation to be in between the two lines representing these observations in the same category, is the lower probability that the next observation belongs to that same category as well. For the upper probability, we consider all possible configurations of segments on the probability wheel, which are consistent with the observations and their corresponding lines on the wheel. The upper probability is then the maximum amount of probability, under $\textcircled{A}_{(n)}$ and these data and configurations, that can be assigned to the segments corresponding to the event of interest.

The assumption that each observation category is represented by a single segment on the probability wheel is crucial to the imprecision in the lower and upper probabilities, and is essential as without this assumption the CA model would lead to vacuous lower and upper probabilities for all non-trivial events.

2.2 Inference for an unknown number of categories

Our inferences in this paper are restricted to a single future observation, which is assumed to be exchangeable with the n observations so far. We will refer to such a future observation as the 'next observation', and will denote it by Y_{n+1}. We will assume that each observation can be assigned to a category with certainty, but we do not require these categories to be defined prior to the observations. We assume that available data consist of n_j observations in category c_j, for $j = 1, \dots, k$, with $\sum_{j=1}^{k} n_j = n$. If the categories are defined upon observation, we have that $n_j \geq 1$, and hence that $1 \leq k \leq n$. We could include further specifically defined categories to our data description, to which no observations belong, but doing so will not influence any of our inferences (as is easily confirmed), so we will not consider this possibility further. For the general setting with unknown total number of possible categories, we must include notation for new, as yet unseen, categories. We distinguish between Defined New categories, of which we need to take the possibility of having several different such categories into account, denoted by DN_i for $i = 1, \dots, l$ for $l \geq 1$, and the possibility that the next observation belongs to any not yet observed category (including categories DN_i), which we describe as an Unobserved New outcome and denote as $Y_{n+1} = UN$. By allowing $l \geq 0$ and $0 \leq r \leq k$ in this notation, we can define two types of events that comprise the most generally formulated events that need to be considered for Y_{n+1} in our multinomial setting. These two general events are

$$Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \backslash \bigcup_{i=1}^{l} DN_i \tag{2}$$

and

$$Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup \bigcup_{i=1}^{l} DN_i \tag{3}$$

Excluding one or more defined new categories in the event of interest, as in (2), can only affect our inferences for events including UN.

The general CA model results for nonparametric predictive inference for the next observation, Y_{n+1}, based on multinomial data, with complete absence of knowledge on the number of possible categories apart from the information provided by $n > 0$ observations, and based on $\textcircled{A}_{(n)}$ and the probability wheel model representation, were presented in [8]. For the first of the general events, the lower probability[3] is

$$\underline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \backslash \bigcup_{i=1}^{l} DN_i) = \begin{cases} \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} - r\right), \text{ for } k \geq 2r \\ \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} - r + \max(2r-k-l,0)\right), \\ \qquad\qquad \text{for } r \leq k \leq 2r \end{cases} \quad (4)$$

and the corresponding upper probability is

$$\overline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \backslash \bigcup_{i=1}^{l} DN_i) = \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} + k - r\right) \quad (5)$$

For the second of these general events, the lower probability is

$$\underline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup \bigcup_{i=1}^{l} DN_i) = \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} - r\right) \quad (6)$$

and the corresponding upper probability is

$$\overline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup \bigcup_{i=1}^{l} DN_i) = \begin{cases} \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} + k - r\right), \text{ for } r \leq k \leq 2r, \\ \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} + r + \min(k-2r,l)\right), \\ \qquad\qquad \text{for } k \geq 2r \end{cases} \quad (7)$$

2.3 Inference with a known number of possible categories

If we assume, for the same multinomial setting, that there is a known number of possible categories, denoted by K, then this extra assumption has an effect on the lower and upper probabilities in the CA model [9]. We restrict attention to $K \geq 3$, as for the binomial situation with $K = 2$ NPI can be based on an assumed data representation on a line, as presented by Coolen [6], which leads to slightly less imprecision than a representation on a circle as in this paper. We can now denote the $K \geq 3$ possible categories by $C_1, \ldots, C_K$, even if their precise definition might only be possible following observations. Without loss of generality, we assume that the first k of these, $C_1, \ldots, C_k$ for $1 \leq k \leq K$, have already been observed and the last $K-k$, $C_{k+1}, \ldots, C_K$ have not yet been observed. Let n_j be the number of observations in C_j, so $n_j \geq 1$ for $j \in \{1, \ldots, k\}$ and $n_j = 0$ for $j \in \{k+1, \ldots, K\}$, and $n = \sum_{j=1}^{k} n_j$. The two general events of interest introduced before, when K was not known, are now reduced to a single general event,

$$Y_{n+1} \in \bigcup_{j \in J} C_j \quad (8)$$

with $J \subseteq \{1, \ldots, K\}$, but except where mentioned explicitly we exclude the trivial events $J = \emptyset$ and $J = \{1, \ldots, K\}$ from our considerations. Let $OJ = J \cap \{1, \ldots, k\}$ denote the index-set for the categories in the event of interest that have already been observed, and $UJ = J \cap \{k+1, \ldots, K\}$ the corresponding index-set for the categories in the event of interest that have not yet been observed. Let r be the number of elements of OJ and l the number of elements of UJ, so $0 \leq r \leq k$ and $0 \leq l \leq K-k$. This implies that $k-r$ observed categories and $K-k-l$ unobserved categories are not included in the event of interest.

The lower and upper probabilities for event (8), according to the CA model with K known, are [9]

$$\underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n}\left(\sum_{j \in OJ} n_j - r + \max(2r+l-K,0)\right) \quad (9)$$

and

$$\overline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n}\left(\sum_{j \in OJ} n_j - r + \min(2r+l,k)\right) \quad (10)$$

For the two trivial events, the NPI-based lower and upper probabilities are obvious. If $J = \{1, \ldots, K\}$, the upper probability of event (8) is equal to 1, in line with (10), and also the lower probability (9) is trivially defined as 1, which is fully in line with the probability wheel representation which underlies the

[3] All probabilities in this paper are predictive given the first n observations, we do not explicitly mention the dependence on the first n observations in the notation.

CA model. Similarly, if $J = \emptyset$, the lower probability of event (8) is equal to 0, in line with (9), and the upper probability (10) is defined as 0. In our further discussion, we will not explicitly mention these trivial events anymore. At the end of this section, we briefly illustrate (9) and (10) via an example, which will be generalized to include sub-categories in Section 3.

2.4 Fundamental properties of the inferences

To derive all the above lower and upper probabilities, we consider all possible configurations σ on the probability wheel, apply $\textcircled{A}_{(n)}$ to each of these to obtain lower and upper predictive probabilities $\underline{P}_{\sigma}(\cdot)$ and $\overline{P}_{\sigma}(\cdot)$, and then take the lower and upper envelope with respect to the set Σ of all configurations [8, 9]. In case of known K, there are fewer configurations on the probability wheel possible for some events of interest, but never more, than when no maximum number of possible categories is known or assumed, hence lower and upper probabilities can be less imprecise if K is known then for the corresponding event in the more general case, but they can never be more imprecise. Actually, such lower and upper probabilities are nested in the logical manner, as the optimization procedures to derive the lower and upper probabilities also take all configurations into account corresponding to known K in the case of an unknown number of categories. All these lower and upper probabilities satisfy a number of important and attractive properties [8, 9]: **(a)** they satisfy the conjugacy property in interval probability theory, and beyond that they are, by applying arguments along the line of [1] (see also [12]), actually F-probability in the sense of Weichselberger [24, 25] and they are coherent in the sense of Walley [22]; **(b)** Corresponding lower and upper probabilities always contain the empirical probability for the event of interest; **(c)** in the limiting situation with $(n \to \infty)$, corresponding lower and upper probabilities become identical. The properties named in (a) imply that the CA model provides sound interval-probabilistic statistical inferences, with strong internal consistency properties. Properties (b) and (c) ensure that these inferences are sensible from classical statistical ('frequency') perspective. Convenient expressions to calculate the lower and upper expectations as simply weighted sums instead of solutions to linear optimization problems are presented in [9].

Properties of the CA model have been discussed in detail before, both for the situations with an unknown total number of possible categories [8] and with at most K possible categories [9], in those papers the resulting inferences were also compared with corresponding inferences based on Walley's Imprecise Dirichlet Model (IDM) [23]. We advocate in particular the fact that the inferences from the CA model do not generally satisfy Walley's 'Representation Invariance Principle' [23], as there are particular situations where for example the number of different categories observed so far would logically have an impact on predictive inference for some events of interest, including events involving categories that have not yet been observed. The CA model provides an attractive alternative to the IDM, and is particularly different on details which were remarked upon by many discussants of Walley's paper [23]. Of course, in situations with substantial data available and a limited number of categories, inferences based on the CA model and the IDM are very similar, in the limit these all agree with empirical probabilities converging to the underlying probabilities (derived from the sizes of the segments). An obvious advantage of the IDM is the fact that it is directly based on a parametric model, with a class of priors used in a similar manner as common in robust Bayesian methods [3]. This implies that inferences can be both in terms of the model parameters and of the future observations, the latter via the class of posterior predictive distributions corresponding to the class of priors chosen [4]. However, as many inferences can be formulated predictively in an attractive and natural manner [7, 16], this apparent advantage of the IDM over the CA model does not hinder applicability of the latter too much.

2.5 An illustrative example

Example 1 briefly illustrates multinomial NPI with a known number of categories, hence formulae (9) and (10) are used.

Example 1.
Suppose that there are $K = 6$ possible categories, namely Blue, Red, Yellow, Green, White, Other, henceforth also indicated by their first letter. Suppose that $n = 9$ observations are available, with the following numbers per category: $B-3$, $R-1$, $Y-2$, $G-3$, $W-0$, $O-0$. We illustrate NPI for the 10th observation, Y_{10}, under the usual assumptions for NPI for multinomial data, as discussed in this section and in more detail in [8, 9]. Some lower and upper probabilities for the events concerning Y_{10} are given in Table 1, it is easy to check that these results illustrate (9) and (10).

$Y_{10} \in \{\cdot\}$	$[\underline{P}, \overline{P}]$
B	$[2/9, 4/9]$
B, R	$[2/9, 6/9]$
B, R, Y	$[3/9, 7/9]$
B, R, Y, G	$[7/9, 1]$
B, R, Y, G, W	$[8/9, 1]$

Table 1. Some lower and upper probabilities (Ex. 1)

For the illustration of our inferences by this example it is helpful to look at the increasing sequence of events described in Table 1. As a consequence of Theorem 2 in [9], where two-monotonicity of $\underline{P}(\cdot)$ was proven, there exists a "least favorable configuration" producing all the lower probabilities of the elements of the sequence as well as a "most favorable configuration" related to all the upper probabilities. For the lower probability note that the probability assigned to a colour that has been observed $n_j - 1$ times is at least $(n_j - 1)/n$. This already gives the whole contribution of the colour to the lower probability as long as there are enough colours not in the event of interest to separate the segments, in order to avoid having to attribute further probability mass $1/n$ to the segment connecting two neighbouring colours in the event of interest. Consequently, we obtain the lower probabilities by the following configuration, where B and R are separated by O and R and Y by W, while Y and G can not be separated anymore, and so additional masses contribute to the lower probability of the event $\{B, R, Y, G\}$, i.e. its lower probability exceeds $\sum_{j \in OJ} n_j - r$.

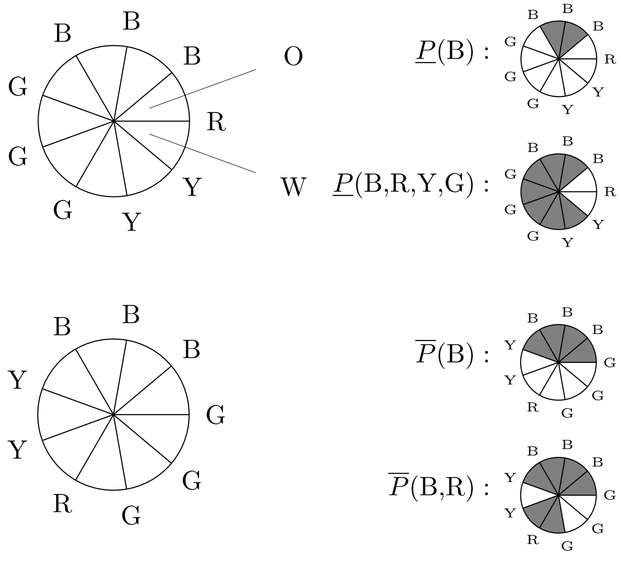

Figure 1. Configurations leading to the lower and upper probabilities in Example 1

Similar arguments apply to the derivation of the upper probability. The main difference is that we now want to assign as much probability mass as possible to the colours in the event of interest, and so we assume that not yet observed colours that are also not in the event of interest do not occur on the probability wheel at all. Again we separate the colours in the event of interest as far as possible, but now with the aim to add probability mass $1/n$ as much as possible. This leads to the configuration at the bottom of Figure 1.

3 Sub-categories

3.1 The modelling of sub-categories

In this section we present the basic principle for dealing with sub-categories in the CA model, and we illustrate this via a basic example. The highest level of categories, in line with the categories as presented in Section 2, will occasionally be referred to as 'main categories', where it is relevant to distinguish these from sub-categories. It is assumed that a main category might be divided into several sub-categories, in such a way that sub-categories are not overlapping and that each sub-category is only related to a single main category. We will assume that each observation belongs to a single main category, and where applicable also to a single sub-category. Such a setting with sub-categories appears, for example, in hierarchical classifications (e.g.[17]). As for the basic CA model (Section 2), both variations with known and unknown total number of possible sub-categories per main category can be dealt with, we restrict our discussion mostly to situations where these numbers are known. We briefly discuss some generalizations in Section 4.

The general principle for dealing with sub-categories is as follows. Each main category is assumed to be represented, in the CA model [8, 9], by a single segment on a probability wheel, with no information about the configuration of all segments representing observed and other relevant categories (those that play a role in predictions). Lower and upper predictive probabilities for the next observation, based on the CA model, are computed by combining this assumed representation with the appropriate $Ⓐ_{(n)}$ assumption, and via minimization and maximization, respectively, over all configurations that are possible for the given data and categories considered. Suppose now that a particular category (e.g. 'Blue') is divided into sub-categories (e.g. 'Light Blue', 'Dark Blue', 'Other Blue'), then this is included in the probability wheel representation underlying the CA model by assuming that the single segment representing the main category is divided into sub-segments, where it is again assumed that each sub-category is represented by a single sub-segment. We have no knowledge, and wish to make no assumptions, about any particular ordering of such sub-segments, hence for events involving one or more sub-segments, the predictive lower and upper probabilities are again derived via the usual $Ⓐ_{(n)}$ assumption, which remains unaffected by the appearance of sub-categories, and minimization and maximization over all possible configurations, now also considering all possible configurations of sub-categories within each corresponding main category. Of course, if one has at least two main categories for which observations are available, then the segments representing the sub-categories of one main category do not form the full circle of the probability wheel, so the combinatorical arguments and computations involved with the sub-categories differ slightly from those for the main categories, yet the principle is straightforward. Note that, if one were to add a new single 'higher-level' category, with the main categories all considered to be sub-categories of this higher-level category, than this makes no difference to the CA model inferences as via the optimisation over all possible configurations that single 'higher-level' category would have no effect whatsoever. It is again possible, as for the events in Section 2 which only considered one level of categories, to derive general expressions for lower and upper probabilities for general events, these have yet to be derived and hence they will be presented at a later stage. It is easily seen that the key properties for NPI for multinomial data discussed in Section 2, including the F-probability property and coherence, can again be rigorously proven in the same way as was done in [9], when sub-categories are included in the model.

We illustrate the general and natural manner in which the CA model can deal with sub-categories in Example 2. For ease of presentation, we restrict attention in this example to a situation with known number K of possible categories [9], as we focus on inferences involving sub-categories. For the case with an unknown number of possible categories, the manner in which the CA model enables sub-categories to be taken into account is identical. We also mostly consider only the case of a known number of sub-categories, this can be generalized to an unknown number of sub-categories in a manner that logically combines the presented way for dealing with sub-categories and the general method for dealing with an unknown total number of categories [8]. As a final restriction to keep presentation at a basic level, we only consider sub-categories of a single main category, of course sub-categories of other main categories are dealt with in the same manner, and one can, for example, generally also consider predictive inference for events involving sub-categories of different main categories. Detailed general results for all such situations will be presented elsewhere.

3.2 Example continued

Example 2.
As in Example 1, suppose that there are six possible categories, Blue, Red, Yellow, Green, White, Other, also indicated by their first letter. In addition, let us assume that observations in Blue are further specified in the sub-categories Light Blue (LB), Dark Blue (DB), or Other Blue (OB). Suppose that 9 observations are available, with the following numbers per (sub-)category: $LB-1$, $DB-2$, $OB-0$, $R-1$, $Y-2$, $G-3$, $W-0$, $O-0$. Of course, these data still imply that there are 3 observations in the main category B, so the lower and upper probabilities for the event $Y_{10} = B$ are as before,

$$[\underline{P}, \overline{P}](Y_{10} = B) = [2/9, 4/9]$$

If we consider events such that Y_{10} belongs to a single sub-category, the resulting lower and upper probabilities are no different from what they would have been if these sub-categories had been main categories, as each is still represented by a single segment on the probability wheel. However, for events involving the union of two sub-categories, the possible configurations of all three sub-categories LB, DB, OB within the main category B must be taken into account. For example, the upper probability for the event $Y_{10} \in \{LB, OB\}$ corresponds to the configurations where DB separates LB, OB within the main category B, while it is irrelevant where B is in the overall configuration with regard to the other main categories, as this is only relevant when events involving unions of main categories are considered, or, as we will discuss later, unions of one or more main categories and sub-categories of other main categories. This separation of LB, OB ensures that of the probability masses that have to be in the main category B, namely two probabilities of 1/9 each, only the probability 1/9 between the two lines representing DB observations has to be assigned to DB, and as LB and OB are on the two extreme sides within the category B, they can now be assigned maximum probabilities of 2/9 and 1/9, respectively. Hence, the upper probability for the event $Y_{10} \in \{LB, OB\}$ is 3/9. The lower probability for this event is 0, as it is easily seen that it is possible (for several configurations) that no actual segment of the probability wheel as created by the data and reflecting the probability masses as assigned by $\textcircled{A}_{(n)}$ in the CA model, [8, 9] must belong to either LB or OB. With similar derivations the lower and upper probabilities presented in Table 2 are derived (see also the example in Figure 2).

The last event in Table 2 is, of course, identical to $Y_{10} = B$. If we had introduced multiple 'Other Blue' sub-categories (OB_i), with no observations for each as yet, then the upper probability that Y_{10} was in any of such sub-category would be equal to 2/9 in case of two such OB_i, and 3/9 in case of three or more of such OB_i, the latter case in agreement with the possible use of UN (see Section 2 and [8]) for such sub-categories if we had not made any assumptions on the number of sub-categories of Blue.

$Y_{10} \in \{\cdot\}$	$[\underline{P}, \overline{P}]$
LB	$[0, 2/9]$
DB	$[1/9, 3/9]$
OB	$[0, 1/9]$
LB, DB	$[1/9, 4/9]$
LB, OB	$[0, 3/9]$
DB, OB	$[1/9, 4/9]$
LB, DB, OB	$[2/9, 4/9]$

Table 2. Some lower and upper probabilities (Ex. 2)

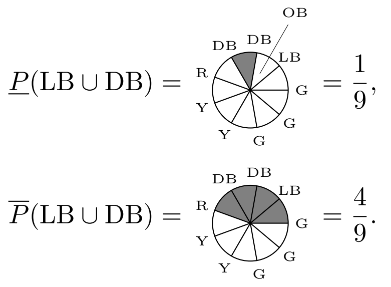
$\underline{P}(\mathrm{LB} \cup \mathrm{DB}) =$
OB
DB
DB
LB
R
G
Y
G
Y
G
$= \frac{1}{9},$
$\overline{P}(\mathrm{LB} \cup \mathrm{DB}) =$
DB
DB
LB
R
G
Y
G
Y
G
$= \frac{4}{9}.$

Figure 2. On the lower and upper probability of $LB \cup DB$

Let us also briefly consider unions of these sub-categories with other main categories. It should be emphasized that considering the information at sub-category level within the main category Blue, or just at the main category level, has no effect whatsoever on events which do not involve B or any of its sub-categories, due to the fact that all considerations of more detailed configurations to deal with the sub-categories only took account of the segment representing Blue, and did not affect the configurations at main categories level. Lower and upper probabilities for events such as $Y_{10} \in \{DB, Y\}$ are derived as usual, as a single sub-category is involved they are identical to corresponding lower and upper probabilities that would correspond to the situation with DB considered as a main category, so this event has lower probability 2/9 and upper probability 6/9. If more sub-categories of the same main category are included in the event of interest, then the same considerations as discussed above must be taken into account, so all configurations of the sub-categories within the main category must be included in the analysis. For such events including other main categories, however, we must combine this with the configurations at the main categories level, which again becomes mainly important in the case of a known total number of main categories and events involving more than half of these [9]. For example, the lower probability for the event $Y_{10} \in \{LB, OB, R, G, W, O\}$ is equal to 3/9, as only the main category Y and sub-category DB are not included, and as long as Y and DB are not next to each other in the configuration, they can both get a maximum of 3 segments assigned out of the 9 in which the observations have divided the probability wheel, where each such segment represents predictive probability 1/9. The upper probability for this event is 7/9, as all but two segments can be assigned to all (sub-)categories in this event of interest. Of course, this also illustrates the conjugacy property with regard to the complementary event $Y_{10} \in \{DB, Y\}$ considered above. For this event involving 2 of the 3 specified sub-categories of B, and 4 of the 5 main categories other than B, clearly there are no possible configurations with all these 6 (sub-)categories included in the event separated from each other by other categories with each at least one observation in them. If this last situation were the case, the upper probability would have been identical to the sum of the upper probabilities of the events $Y = X$ with $X \in \{LB, OB, R, G, W, O\}$ [9].

To emphasize the difference between sub-categories and main categories, let us compare the event $Y_{10} \in \{LB, DB\}$, which has lower and upper probabilities 1/9 and 4/9, with the event $Y_{10} \in \{R, Y\}$. For both sets of (sub-)categories in these events, we have one (sub-)category with a single observation, and one with two observations. However, the lower and upper probabilities for the event $Y_{10} \in \{R, Y\}$ are equal to 1/9 and 5/9, so this upper probability is larger than that for $Y_{10} \in \{LB, DB\}$. This results from the fact that the main categories R and Y can be fully separated, in the configurations for the probability wheel representation, by categories with positive numbers of observations in them, whereas the sub-categories LB and DB can only be separated by OB, in which there are no observations. If one of the observations in G had actually been in OB, then both these events considered here would have had the same upper probability 5/9.

It will be clear from this example that the CA model, as before, does not satisfy Walley's 'Representation Invariance Principle' (RIP) [23], a fact which we have commented on in detail before [8, 9], and which we do perceive as an advantage of our model. One could argue, however, that the fact that, in the CA model, it does not matter whether one uses information at main category level, or at sub-category level, as long as this category is not involved in the event of interest, is very close in nature to the underlying idea of Walley's RIP.

4 Concluding remarks

In this paper we have reviewed the CA model as presented so far [8, 9], and we have outlined the general manner in which the CA model can deal with data at sub-category level, to get consistent inferences at both main and sub-category levels. Detailed expressions for lower and upper probabilities, for general events in a variety of situations with regard to assumed knowledge of numbers of (sub-)categories will be presented elsewhere, but all follow the basic concept outlined in Section 3 and illustrated in Example 2. This generalization of the CA model is of great practical use, as interest is often explicitly at sub-category levels, with potentially even more layers of sub-categories playing a role. As long as such different layers are representable by tree structures, the same approach as outlined here can be used, guaranteeing strong internal consistency of inferences at varying levels due to the use of the probability wheel representation. It remains important here that no actual ordering of (sub-)categories is known. If one wishes to use a multinomial approach with categories ordered, as for example Coolen [5] did for lifetime data on the basis of Walley's IDM, then the CA model with the probability wheel representation might not be suitable. In particular if one models time categories, with a natural one-dimensional ordering, the general framework of NPI offers more suitable modelling opportunities, as Coolen and Yan [10] presented for grouped lifetime data, using another variation of Hill's $A_{(n)}$ for dealing with right-censored data [11].

Throughout this paper, and in [8, 9], we assume to have perfect information on each observation, that is we know with certainty which unique (sub-)category it belongs to. If only partial information is available, in the sense that it is only known for a particular observation to belong to a subset of (sub-)categories [21, 27], then the CA model is easily adapted to deal with such information in a consistent manner, taking all possibilities of the values of that particular observation into account and again optimizing over all possible corresponding configurations of the observations on the probability wheel. However, all such generalizations make it harder to derive general expressions for the lower and upper probabilities for events of interest, as the combinatorial problems in deriving analytic solutions of the optimization processes involved become ever more complex.

In the CA model, as in NPI in general [2, 7], updating in the light of new observations is straightforward, as simply new lower and upper probabilities are calculated on the basis of the entire data set. Conditioning, however, is more complex [2], where conditioning is understood as taking additional information into account on the particular random quantity of interest, in contrast to information in the form of further observed exchangeable random quantities in updating. Generalization of the classical, precise probabilistic, concept of conditioning is acknowledged to be a complex issue in theory of lower and upper probability [24, 26], and this is not any different in conditioning within the CA model. For example, suppose that for the situation in Example 2, one learns that Y_{10} is actually Blue, but that one then is interested in which of the three specified sub-categories it belongs to. Following the basics of the NPI approach, and of the CA model, a correct way of arguing is that of the nine observations so far, only three can still be assumed to satisfy the post-data exchangeability assumption that is key for any inference based on $A_{(n)}$ and its variations such as $\textcircled{A}_{(n)}$, namely the three already observed Blue outcomes, of which one was LB and two were DB, with OB as only other sub-category assumed. Hence, instead of considering Y_{10} with a post-data exchangeability assumption with 9 available observations, one should now redefine the random quantity of interest as, say, $\tilde{Y}_4$, with 3 observations available, and (if deemed appropriate) one can use $\textcircled{A}_{(n)}$ with the three sub-categories now functioning as main categories, in which case the lower and upper probabilities for events involving $\tilde{Y}_4$ are easily derived using (9) and (10). Generally, the lower and upper probabilities for $\tilde{Y}_4$ derived in this manner are not proportional to those for the corresponding events involving Y_{10} and based on all 9 observations, before taking the information $Y_{10} = B$ into account. Although this is not a surprise due to the complex general nature of conditional lower and upper probabilities, detailed study of properties of such conditioning within the CA model is an important topic for future research.

Acknowledgements

We are grateful to referees for helpful suggestions on presentation. We were extremely impressed by the efforts of one referee who was kind enough to provide Latex code based on the tikz-package for the creation of the probability wheel plots used in this paper.

References

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

[2] T. Augustin and F.P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124: 251-272, 2004.

[3] J.O. Berger. Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25: 303-328, 1990.

[4] J.M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39: 123-150, 2005.

[5] F.P.A. Coolen. An imprecise Dirichlet model for Bayesian analysis of failure data including right-censored observations. *Reliability Engineering and System Safety*, 56: 61-68, 1997.

[6] F.P.A. Coolen. Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36: 349-357, 1998.

[7] F.P.A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15: 21-47, 2006.

[8] F.P.A. Coolen and T. Augustin. Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. In: *Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications*, F.G. Cozman, R. Nau and T. Seidenfeld (Eds), pp. 125-134, 2005.

[9] F.P.A. Coolen and T. Augustin. A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. In revision for *International Journal of Approximate Reasoning*. SFB-Disc. Paper 489: `http://www.stat.uni-muenchen.de/sfb386/`

[10] F.P.A. Coolen and K.J. Yan. Nonparametric predictive inference for grouped lifetime data. *Reliability Engineering and System Safety*, 80: 243-252, 2003.

[11] F.P.A. Coolen and K.J. Yan. Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126: 25-54, 2004.

[12] G. de Cooman, E. Miranda, I. Couso, Lower previsions induced by multi-valued mappings, Journal of Statistical Planning and Inference 133: 173-197, 2004.

[13] A.P. Dempster. On direct probabilities. *Journal of the Royal Statistical Society, Series B*, 25: 100-110, 1963.

[14] B. De Finetti. *Theory of Probability*. Wiley, Chichester, 1974.

[15] S. French and D. Rios Insua. *Statistical Decision Theory*. Arnold, 2000.

[16] S. Geisser. *Predictive Inference: an Introduction*. Chapman and Hall, New York, 1993.

[17] A.D. Gordon. *Classification* (2nd Edition). Chapman and Hall, Boca Raton, 1999.

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

[19] B.M. Hill. De Finetti's Theorem, Induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In: *Bayesian Statistics 3*, Bernardo et al. (eds.), Oxford University Press, pp. 211-241, 1988.

[20] J.F. Lawless and M. Fredette. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92: 529-542, 2005.

[21] L.V. Utkin, T. Augustin. Decision making under incomplete data using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44: 322-338, 2007.

[22] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

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

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

[25] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg, 2001.

[26] K. Weichselberger, T. Augustin. On the competition and symbiosis of two concepts of conditional interval probability. In: *Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, J.M. Bernard, T. Seidenfeld and M. Zaffalon (Eds), pp. 608-629, 2003.

[27] M. Zaffalon. Conservative rules for predictive inference with incomplete data. In: *Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications*, F.G. Cozman, R. Nau and T. Seidenfeld (Eds), pp. 406-415, 2005.