Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model

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

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

Abstract

A new model for learning from multinomial data has recently been developed, giving predictive inferences in the form of lower and upper probabilities for a future observation. Apart from the past observations, no information on the sample space is assumed, so explicitly no assumptions are made on the number of possible categories. In this paper, we briefly present the general lower and upper probabilities corresponding to this model, and illustrate their properties via two examples taken from Walley's paper [16], which introduced the imprecise Dirichlet model (IDM). As our approach is nonparametric, its applicability is more restricted. However, our inferences do not suffer from some disadvantages of the IDM.

Keywords. Imprecise Dirichlet model, imprecise probabilities, interval probability, multinomial data, nonparametric predictive inference, probability wheel.

1 Introduction

In this paper, we present direct lower and upper probabilities for a future observation, based on observed multinomial data. We start by introducing notation used for such multinomial data and for the events of interest. Then we explain nonparametric predictive inference (NPI) [2], as based on Hill's assumption $A_{(n)}$ [14], and we introduce the variation of this assumption suitable for the assumed model on which our predictive lower and upper probabilities are based. In Section 2, we present the general lower and upper probabilities for a future observation in our multinomial setting, and we give a brief description of the assumed model on which these lower and upper probabilities are based. In Section 3, we present two examples to illustrate our lower and upper probabilities. In Section 4, we compare our approach with Walley's Imprecise Dirichlet Model [16], which has attracted considerable attention (see [3] for a survey of work based on the IDM). We end this paper with a brief discussion of some related aspects, these will be discussed in far more detail, together with a detailed justification of the results presented here, elsewhere [6].

1.1 Multinomial data and notation

In a standard multinomial setting, observations belong to categories, with no natural relationships or orderings between these categories. 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. Throughout this paper, we assume that available data consist of n_j observations in category c_j, for $j = 1, \ldots, k$, 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. In this paper, lower and upper probabilities based on such data do not normally have the data representation on which they are based explicitly mentioned in the notation, but in the examples in Section 3, where we consider different representations of the data, we will include these explicitly in the notation.

Our inferences in this paper are restricted to a single future observation, which is assumed to be exchangeable with the n observations so far, under the assumed model as discussed in Section 2. We will refer to such a future observation as the 'next observation', and will denote it by Y_{n+1}. For our general results, presented in Section 2, and consisting of lower and upper probabilities for all possible outcomes for the next observation, we must include notation for new, as yet unseen, categories. We need to distinguish between *D*efined *N*ew categories, of which we need to take the possibility of having several different such categories into

account, denoted by DN_i for $i = 1, \ldots, l$ 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 the notation introduced above, 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

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

and

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

We wish to emphasize that excluding one or more defined new categories in the event of interest, as in (1), can only affect our inferences for events including UN.

1.2 Nonparametric predictive inference and interval probability

Hill [14] introduced the assumption $A_{(n)}$ as a basis for predictive inference in case of real-valued observations. In his setting, 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 [10], which provides direct posterior predictive probabilities [11]. Hill [14] 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 that nonparametric predictive inference (NPI) based only on $A_{(n)}$ has strong consistency properties in the theory of interval probability [15, 18, 19].

For circular data $A_{(n)}$ is not suitable, as the data are not represented on the real-line. A straightforward variation, again linked to exchangeability of $n+1$ observations, is the assumption that we call *circular*-$A_{(n)}$ and denote by $\mathcal{A}_{(n)}$:

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 $\mathcal{A}_{(n)}$ is that a future observation X_{n+1} falls into each of these n intervals with equal probability, so

$$P(X_{n+1} \in I_j) = \frac{1}{n}, \text{ for } j = 1, \ldots, n. \tag{3}$$

Notice that neither the units of the circular data, nor the chosen 0-point on the circle, are relevant here. $\mathcal{A}_{(n)}$ is clearly 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. Hence, nonparametric predictive inference based on $\mathcal{A}_{(n)}$ has the same consistency properties as such inference based on $A_{(n)}$ [2].

In this paper, we use $\mathcal{A}_{(n)}$ combined with an assumed underlying representation of multinomial data as outcomes of spinning a probability wheel (see Section 2). As we wish not to make further assumptions about the probability mass $1/n$ per interval I_j, our predictive inferences are in the form of interval probabilities [2, 15, 18, 19], 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 $\mathcal{A}_{(n)}$, according to De Finetti's fundamental theorem of probability [10], and the upper probability is the minimum upper bound consistent in this way. From a subjective point of view as advocated by Walley [15], these can also be interpreted as maximum buying and minimum selling prices, respectively, for which one judges gambles on the event A to be desirable. The predictive lower and upper probabilities presented in this paper lead to F-probability in the interval probability theory of Weichselberger [2, 18, 19], see the Appendix for more details. This proves that these predictive interval probabilities, based on a particular data representation, are internally consistent in a very strong sense: The resulting limits are in complete accordance with a non-empty set of classical ('precise') σ-additive probabilities, and so the bounds make perfectly use of the available information; they are neither too wide nor do they add unjustified additional assumptions to our inferences. Additionally, the F-probability property also implies coherence, and avoiding sure loss, in Walley's sense [15]. As a consequence, our bounds are also perfectly rational from the behavioral point of view. Important properties of F-probability are that the lower and upper probabilities contain a classical probability, that the lower (upper) probability is superadditive (subadditive), and that $\underline{P}(A) = 1 - \overline{P}(\bar{A})$, where $\bar{A}$ is the

complementary event to A.

2 General results

2.1 Predictive lower and upper probabilities

We now present the general 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 $\mathcal{A}_{(n)}$ and the probability wheel model representation which we discuss in some more detail in Section 2.2 and in [6].

For the first of the general events introduced in Section 1, the lower probability is

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

and the corresponding upper probability is

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

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

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

It is easy to confirm that these lower and upper probabilities satisfy $\underline{P}(A) = 1 - \overline{P}(\bar{A})$ for all events A considered for the next observation, for a chosen data representation. The classical probability defined as the relative frequency of A in the data, is always bounded by these $\underline{P}(A)$ and $\overline{P}(A)$. In the Appendix we justify our claim that these lower and upper probabilities, based on a particular data set and data representation, lead to F-probability. In [6] we will present a detailed study of the properties of these lower and upper probabilities in the theory of interval probability [18, 19], including attention to conditioning and updating (cf. [2]). It is of interest to consider the events for which our lower probabilities are equal to 0. For both events (1) and (2), this only occurs if $n_{j_s} = 1$ for all $s = 1, \ldots, r$, but for event (1) a further condition is required if $k < 2r$, in which case $\underline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \setminus \bigcup_{i=1}^{l} DN_i) = 0$ only occurs if also $l \geq 2r - k$.

2.2 Underlying model

The predictive lower and upper probabilities presented in Section 2.1 are based on an underlying assumed model, ensuring that they not only make sense for one specific set of data (and a chosen corresponding data representation), which they do being F-probability and due to the fact that they bound the observed relative frequencies, but are also consistent if more observations are added to the data. Such considerations will be discussed in detail in [6], together with the underlying model and the principles leading to, and detailed justification of, the above presented lower and upper probabilities. Here, we give a brief summary of the key aspects of this model and justification.

The model underlying our nonparametric predictive lower and upper probabilities (4)-(7) 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 [12] 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. In this paper, we assume that the only information available to us is a

finite number of exchangeable observations. Such observations define observation categories in the sense that, on each observation, we know with certainty if it is from a category not previously observed, or to which previously observed category it belongs. We assume to have no further knowledge about the total number of categories on the probability wheel, which can conceptually be 'very large' - as we are only interested in prediction for a finite number of future outcomes (only one in this paper), based on a finite number of observations, we do not need to assume whether or not 'very large' might include 'infinite'. 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 $\mathcal{A}_{(n)}$ once we consider unions of two or more categories, and leads to imprecision of our inferences, in the sense that our lower and upper probabilities are optimal bounds over all possible configurations, including possibly assumed additional unseen categories.

When we combine this concept of a probability wheel, with each observation category represented by a single segment, with the assumption $\mathcal{A}_{(n)}$, 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 in n equally sized slices, with the next observation 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 $\mathcal{A}_{(n)}$, 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 $\mathcal{A}_{(n)}$ and these data and configurations, that can be assigned to the segment(s) corresponding to the event of interest.

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

3 Examples

In this section we illustrate our inferences via two examples, both taken from Walley [16], but with variations to emphasize special features of our method. These examples are also central to the comparison of our method with Walley's IDM in Section 4. Following Walley, we use colours as different categories, but we should emphasize that while this helps presentation, one should be careful not to link different categories intuitively, or for example take any natural ordering or upper bound on the number of colours into consideration.

Example 1.
Walley [16] advocates the IDM, in particular its property that IDM-based inferences do not depend on choice of the sample space. He discusses the following example to motivate this property. Assuming three observations, Walley considers the event that the next observation is Red or Yellow. The sample spaces considered are:
(a) {Red or Yellow, Other observed colours}, with the data reported as '1 Red or Yellow, 2 Other observed';
(b) {Red, Yellow, Other observed}, with the data reported as '0 Red, 1 Yellow, 2 Other observed';
(c) {Red or Yellow, Blue, Green, White, Other observed}, with the data reported as '1 Red or Yellow, 1 Blue, 0 Green, 1 White, 0 Other observed'.
For our inferences, the most detailed data representation corresponding to these observations can be denoted as $D_d = (Y : 1; B : 1; W : 1)$, where each colour is represented by its first letter. It is straightforward to see that, in our model, explicit reporting of not observed data categories does not influence our inferences, as they do not affect the data representation in the underlying probability wheel model. This data representation D_d was not explicitly used by Walley. Relating to the three sample spaces used by Walley, we consider the following data representations:

(a) $D_a = (RY : 1; O : 2)$;
(b) $D_b = (Y : 1; O : 2)$;
(c) $D_c = (RY : 1; B : 1; W : 1)$;
(d) $D_d = (Y : 1; B : 1; W : 1)$.

RY denotes the category 'Red or Yellow', and O the category 'Other observed', that is here 'Observed and not Red or Yellow'. Clearly, O forms a single category here - it may be the case (as (c) suggests) that the 2 observations in this category could well be distinguished, but once recorded to belong to this category, such information is not used further and hence should be considered not to be available.

Our NPI-based lower and upper probabilities for the event that the fourth observation is Red or Yellow are given below, for the four data representations above. Depending on the data representation, so on the definition of the observation categories, this event is either denoted as $Y_4 = RY$ or as $Y_4 \in \{R, Y\}$.

(a) $[\underline{P}, \overline{P}](Y_4 = RY|D_a) = [0, 2/3]$;
(b) $[\underline{P}, \overline{P}](Y_4 \in \{R, Y\}|D_b) = [0, 2/3]$;
(c) $[\underline{P}, \overline{P}](Y_4 = RY|D_c) = [0, 2/3]$;
(d) $[\underline{P}, \overline{P}](Y_4 \in \{R, Y\}|D_d) = [0, 1]$.

In contrast, Walley's IDM leads to the same lower and upper probabilities in these 4 cases, $\frac{1}{3+s}$ and $\frac{1+s}{3+s}$, respectively, with $s > 0$ a constant which Walley emphasizes should be chosen independently of the data representation, and for which Walley suggests that, for example, values $s = 1$ or $s = 2$ may be suitable. Our method takes the different data representations explicitly into account, let us consider this in detail. We must emphasize that, although we could compare these lower and upper probabilities corresponding to different data representations from the overall perspective of knowing D_d, this does not logically imply that all these lower and upper probabilities should be identical. Clearly, once data are recorded in defined categories, from these recorded data alone one cannot deduce the more detailed information about the observations, for example if we only have data represented as D_b, we cannot deduce the more detailed representation D_d from this anymore.

Our lower probabilities in cases (a)-(d) are all equal to 0. In our model, this is a consequence of only having one observation in the category $\{R, Y\}$ or RY, while our predictive lower probability can only be positive if the category of interest has been observed more than once. Observing a category once does not strongly imply that it can be observed again, unless one wants to add assumptions on the possible number of categories [4].[1] One might argue that it may be more likely that the next observation is an outcome which has been observed once before than a defined new category. We would agree with such an intuition, in our approach this is reflected in different upper probabilities for such events.

Our upper probabilities in cases (a)-(d) vary both due to the representation of the event of interest, so whether or not Red and Yellow are combined into a single category or not, and due to the representation of the other observations in one or two categories. Let us consider these upper probabilities from the perspective of our assumed probability wheel model, and the corresponding assumption $\mathcal{A}_{(n)}$ for the fourth observation. In cases (b) and (d), 'Red' is a different category to 'Yellow', and as there is no 'Red' observation yet, 'Red' is a *D*efined *N*ew category in the event of interest. For case (d), the 3 lines representing the data on the probability wheel all belong to different categories. Without further assumptions on the sizes of the category segments on this wheel, the segment between the B and W lines could be 'Red', and the other two segments have the Y line as one of its boundaries, hence these can be all 'Yellow', hence the probability wheel could, effectively, be all either 'Red' or 'Yellow', apart from the two observed B and W lines. This implies that two of the three observations had been at lines which, if this extreme configuration were the actual probability wheel, would have had probability zero to occur. However, any positive lower probability for B or for W should correspond to a segment of the probability wheel which, in the current representation, must have a positive size, and on the basis of a single observation any such a necessarily positive size can only be based on additional assumptions. For case (b), we have two undistinguished 'Other' observations, represented on our probability wheel by two lines belonging to the same segment. Hence, the size of this 'Other observed' segment in the current representation is at least 1/3, the other two segments can be 'Red' or 'Yellow'. The same reasoning applies to case (a), the only difference being that RY is a single category, but it can still consist of the two segments which are not in between the two lines representing the O observations. For case (c), the area between the lines representing the observations B and W cannot belong to the segment RY, as every category is represented by only a single segment, and it obviously must contain the line representing the RY observation. Hence, both segments which have the RY line as a bound could all belong to the segment RY, giving upper probability 2/3 again.

This example becomes more interesting if we add a fourth observation, which is Red, and consider our NPI-based inferences for the fifth observation. Let us again compare the lower and upper probabilities for the event that the next observation is either Red or Yellow, where we consider the following data representations:

[1] See also the more detailed discussion of this aspect at the end of Section 4.

(e) $D_e = (R:1;\, Y:1;\, B:1;\, W:1)$;

(f) $D_f = (RY:2;\, O:2)$;

(g) $D_g = (R:1;\, Y:1;\, O:2)$;

(h) $D_h = (RY:2;\, B:1;\, W:1)$.

These data representations lead to the following lower and upper probabilities:

(e) $[\underline{P}, \overline{P}](Y_5 \in \{R, Y\}|D_e) = [0, 1]$;

(f) $[\underline{P}, \overline{P}](Y_5 = RY|D_f) = [1/4, 3/4]$;

(g) $[\underline{P}, \overline{P}](Y_5 \in \{R, Y\}|D_g) = [0, 3/4]$;

(h) $[\underline{P}, \overline{P}](Y_5 = RY|D_h) = [1/4, 3/4]$.

The lower probabilities for these events of interest vary. If data are recorded with 'Red' and 'Yellow' belonging to one category, then the fact that two observations have been made in this category causes the lower probability for the next observation to belong to the category RY to become positive. If one uses the combined category RY then the information about the specific colour of these two observations is not taken into account. The differences in the upper probabilities in cases (e)-(g) are easily understood by the same principle with regard to the chosen representation for the other colours observed. For case (h), the segment RY on the probability wheel, in our assumed model with this data representation, cannot contain the area between the lines representing the observations B and W, which represents predictive probability 1/4, but the whole remaining area of the probability wheel could be assigned to RY, leading to the upper probability 3/4. It is interesting to compare the two extreme data representations, (e) and (f). In case (f), the data have so far been recorded as Bernoulli data ('Red or Yellow' - yes or no), whereas in case (e) all 4 observations were distinguished. Case (e) leads to more predictive imprecision for our event of interest, which reflects the logical fact that, if we represent data in more detail, more information would be required to reduce imprecision than if we represent data in less detail. This will be discussed, as a basic principle of inference, in [6]. Related to case (e), so data D_e, it is interesting to consider a possible fifth observation, also in a new category, say $G:1$, with data represented as

$$D_i = (R:1;\, Y:1;\, B:1;\, W:1;\, G:1).$$

The corresponding predictive lower and upper probabilities for the sixth observation are

$$[\underline{P}, \overline{P}](Y_6 \in \{R, Y\}|D_i) = [0, 4/5],$$

illustrating that with all observations belonging to different categories, such upper probabilities are less than one if more than half of all observations belong to categories not in the event of interest. In such a case the corresponding lower probabilities remain zero[2]. We illustrate the inclusion of such not yet observed categories in Example 2, emphasizing the important difference between the next observation belonging to *D*efined *N*ew categories and, more generally, it just being an *U*nobserved *N*ew outcome.

Example 2.
Walley [16] discusses an example with 6 observations, consisting of 1 Red, 3 Blue and 2 Green. He focusses on the probability that the 7th observation is again Red. We also use this example to illustrate our NPI-based method, but consider more events, in particular events including the possibility that the next observation belongs to a new category, where we distinguish between one or more *D*efined *N*ew categories, and any *U*nobserved *N*ew outcome. We also consider some variations to these data, to illustrate important properties of our method. This example will also be referred to in Section 4, for the general comparison between Walley's IDM and our method.

Let us first represent the 6 available observations by

$$D_1 = (R:1;\, B:3;\, G:2),$$

so the data belong to $k = 3$ different observed categories. This leads to

$$[\underline{P}, \overline{P}](Y_7 = R|D_1) = [0, 2/6].$$

When we also take new categories into account, we get

$$[\underline{P}, \overline{P}](Y_7 \in \{R, UN\}|D_1) = [0, 3/6]$$

and also

$$[\underline{P}, \overline{P}](Y_7 \in \{R, DN\}|D_1) = [0, 3/6],$$

where the upper probabilities are identical as there is only one segment of the probability wheel available, in the current representation corresponding to these upper probabilities, which can be assigned to a new category, namely the segment bounded by one Blue and one Green line. Hence, based on D_1 we also get the same lower and upper probabilities if we include more than one DN in this event of interest.

Let us now suppose that, on reconsideration, one observation was mistakenly classified as Blue, it should have been classified as Purple, and let us use data representation

$$D_2 = (R:1;\, B:2;\, G:2;\, P:1).$$

[2]This is not in conflict with the F-probability requirement that $\underline{P}(A) = 1 - \overline{P}(\bar{A})$, as the complementary event in this case includes not yet observed categories.

This leads to

$$[\underline{P}, \overline{P}](Y_7 = R|D_2) = [0, 2/6],$$

and lower probabilities

$$\underline{P}(Y_7 \in \{R, UN\}|D_2) = 0$$

and

$$\underline{P}(Y_7 \in \{R, DN\}|D_2) = 0,$$

which are the same as for D_1. The upper probabilities for these two events are now

$$\overline{P}(Y_7 \in \{R, UN\}|D_2) = 4/6$$

and

$$\overline{P}(Y_7 \in \{R, DN\}|D_2) = 3/6,$$

while for $l \geq 2$ we have

$$\overline{P}(Y_7 \in \{R\} \cup \bigcup_{i=1}^{l} DN_i|D_2) = 4/6.$$

These upper probabilities correspond to the fact that, under D_2, only one out of the six segments in our current representation is assigned to Blue, and one to Green, with two different segments still available for new categories. Clearly, a single *D*efined *N*ew category can only be assigned to at most one such segment, causing the difference between these upper probabilities. For some events of interest including new categories, also the lower probabilities can be different. For example,

$$\underline{P}(Y_7 \in \{R, B, G, P, UN\}|D_2) = 1,$$

whereas

$$\underline{P}(Y_7 \in \{R, B, G, P, DN\}|D_2) = 2/6,$$

as for the latter event 4 of the 6 segments on the probability wheel, based on data D_2, could be assigned to not yet observed categories other than the one explicitly defined, DN, indeed this lower probability would remain 2/6 for any number of different *D*efined *N*ew categories DN_i included in the event of interest.

Let us now consider the situation that the 6 observations are actually judged all to belong to different categories, with the Blue and Green ones distinguished in light and dark shades, represented by

$$D_3 = (R : 1;\ LB : 1;\ DB : 1;\ LG : 1;\ DG : 1;\ P : 1).$$

We have

$$[\underline{P}, \overline{P}](Y_7 = R|D_3) = [0, 2/6],$$

and the lower probabilities

$$\underline{P}(Y_7 \in \{R, UN\}|D_3) = 0$$

and

$$\underline{P}(Y_7 \in \{R, DN\}|D_3) = 0$$

are the same as for D_1 and D_2. The upper probabilities for these latter two events are now

$$\overline{P}(Y_7 \in \{R, UN\}|D_3) = 1$$

and

$$\overline{P}(Y_7 \in \{R, DN\}|D_3) = 3/6,$$

while

$$\overline{P}(Y_7 \in \{R\} \cup \bigcup_{i=1}^{l} DN_i|D_3) = (2 + l)/6,$$

for $l = 2, 3$, and

$$\overline{P}(Y_7 \in \{R\} \cup \bigcup_{i=1}^{l} DN_i|D_3) = 1,$$

for $l \geq 4$. These upper probabilities correspond logically, by $\underline{P}(A) = 1 - \overline{P}(\bar{A})$, to the lower probabilities of the complementary events, which is particularly clear for the event $Y_7 \in \{R, UN\}$ based on D_3, for which the complementary event has

$$\underline{P}(Y_7 \in \{LB, DB, LG, DG, P\}|D_3) = 0,$$

caused by the fact that none of these categories has been observed more than once. With this data representation, we also have the important difference between

$$[\underline{P}, \overline{P}](Y_7 = UN|D_3) = [0, 1]$$

and

$$[\underline{P}, \overline{P}](Y_7 = DN|D_3) = [0, 1/6].$$

The upper probability for $Y_7 = DN$ is 1/6 for any data representation, but the upper probability for $Y_7 = UN$ depends on the specific data representation, and is less than 1 for data representations with two or more observations belonging to the same category, and it also becomes 1/6 in case all six observations are represented by a single category.

4 Comparison with the Imprecise Dirichlet Model

Walley [16] presented the Imprecise Dirichlet Model (IDM) for inference from multinomial data. According to the IDM, the lower and upper probabilities, based on n observations, for the next observation Y_{n+1} to be in a category C, are

$$[\underline{P}, \overline{P}]_{IDM}(Y_{n+1} \in C) = \left[\frac{n_c}{n + s}, \frac{n_c + s}{n + s}\right], \qquad (8)$$

with n_c the number of observations in C, and s a positive constant, independent of the data. Walley states as important advantage of this model that it satisfies a 'Representation Invariance Principle' (RIP), stating that such lower and upper probabilities should not depend on the sample space in terms of which the event of interest and the data are represented. Our inferences clearly do not satisfy the RIP. We would consider the RIP a reasonably logical principle from the perspective of classical probability, where a precise probability for such inferences should be close to the proportion of observations in the categories specified in the event of interest. However, from the perspective of interval probability theory, it is natural that the difference between corresponding lower and upper probabilities depends on the amount of information available and the data representation. A more detailed data representation allows more detailed inferences, but since it will imply less information on one or more categories, the price for such more detailed inferences can be greater imprecision. This feature of our method is similar in nature to the effects of increasing the number of parameters in a statistical model, which allows the information from the data to be taken into account in more detail, hence leads to improved model fit, but tends to cause loss of predictive power. In our inferences, this latter aspect occurs in the form of more predictive imprecision in case of a more detailed data representation. It is crucial here to emphasize that, once a data representation has been chosen, the corresponding inferences should not be judged from the perspective of actually knowing more details of the data.

The discussants to Walley's paper [16] raised a number of disadvantages for the IDM, and some of these were also mentioned and shared by Walley. These disadvantages of the IDM include the following: (1) The IDM lower probability for the second observation to be equal to the first, is $\frac{1}{1+s}$. Walley suggests to use rather small values of s, in particular $s = 1$ or $s = 2$, both of which lead to an intuitively surprisingly high value for this lower probability. (2) The IDM predictive lower and upper probabilities depend only on the observed frequency of that category and the total number of observations. (3) When considering events including as yet unseen categories, the IDM does not distinguish between defined new categories (DN) and any unobserved new outcome (UN), and, closely related to this; (4) the IDM lower and upper probabilities for the event that the next observation is in an as yet unseen category does not depend on the number of categories seen so far. Our examples in Section 3 show that our lower and upper probabilities do not share these disadvantages. In particular, with regard to (3) and (4), if all n observations belong to the same category, we have

$$[\underline{P}, \overline{P}](Y_{n+1} = UN) = [\underline{P}, \overline{P}](Y_{n+1} = DN) = [0, 1/n],$$

whereas if all n observations belong to different categories, we have

$$[\underline{P}, \overline{P}](Y_{n+1} = UN) = [0, 1]$$

but

$$[\underline{P}, \overline{P}](Y_{n+1} = DN) = [0, 1/n],$$

and, for any data set between these two extremes, the latter event has these same lower and upper probabilities, but the upper probability for the event $Y_{n+1} = UN$ is increasing as function of the number of observed categories, while the lower probability for this event remains zero. The lower and upper probabilities according to the IDM are $[0, \frac{s}{n+s}]$ for both these events, independent of other aspects of the data apart from n. With regard to IDM disadvantage (1), we wish to emphasize again that our corresponding lower probability is 0, as our lower probability for the next observation to belong to a category only becomes positive if that category has been observed twice. The difference between events concerning a category with a single observation, and one not yet observed, is reflected in the upper probabilities. One may argue that, in such a case, the lower probability should also become positive for the event that the next observation will belong to a category with a single observation, but any such a positive value would imply a subjective judgement on the number of different categories that can be observed, which our inferences explicitly avoid. Anyway, if one judges our lower and upper probabilities not to be in accordance with intuition, we hope that mostly one finds that they are perhaps too cautious. For the IDM, several discussants [16] made clear that they felt that the IDM was not cautious enough for several possible events of interest, as mentioned among the disadvantages listed above, when compared to subjective inferences.

We should point out that the IDM also has some important advantages. In particular, as it is a parametric model in the Bayesian framework, it allows a far wider range of inferences than our approach, and it is easily adapted to enable prior judgements to be formally taken into account. In our NPI-based method, inference is necessarily restricted to predictive events, but quite many inferences of practical interest can be naturally formulated in a predictive manner, see for example [7, 9, 13]. A comparison with the currently developed approach based on logical interval probability [20] will be given in [6]. The imprecise Dirichlet-multinomial model, presented by Walley and Bernard [17], gives the same lower and upper probabilities (8)

as the IDM, so it does not provide an alternative solution with regard to the above mentioned disadvantages of the IDM.

5 Discussion

This paper briefly presents lower and upper probabilities derived from a new model to reason with multinomial data, with main focus on comparison with Walley's IDM. We will present our method in detail elsewhere [6], including detailed derivations and justifications of its properties, and discussion of related general interval probabilistic principles for statistical inference. We will also discuss updating and conditioning with these lower and upper probabilities, following our earlier results for general $A_{(n)}$-based nonparametric predictive inference [2]. The results in this paper can, in principle, easily be extended to multiple future observations, by sequential conditioning, but this requires further detailed attention with regard to combinatorial aspects and more general definition of events of interest, in particular as the possibility of two or more different Unobserved New outcomes must be taken into account. It is also possible to use a similar approach in case (an upper bound for) the number of possible categories is known, but this also requires further study, in particular with regard to combinatorial aspects. Including such (assumed) information on the number of categories would reduce imprecision when compared to the lower and upper probabilities presented in this paper.

One may be tempted to apply these inferences to settings where different categories are actually ordered, for example with regard to lifetime inferences. We would not normally recommend this, as more suitable data representations [5, 8] may be available for similar nonparametric predictive inference, which then are likely to lead to less imprecision.

If one has opted for a particular data representation, but is then interested in an event using a subcategory of a category in the data representation, our method may still be applicable in an obvious manner and without further assumptions, but this would lead to greater imprecision than if one had used a more appropriately detailed data representation, and it would require additional knowledge about the definitions of the categories. For example, in relation to Example 1 in Section 3, suppose one had represented the data as $D_a = (RY : 1; O : 2)$ but was actually interested in the event $Y_4 = R$, and that one knew that R could be considered a subcategory of RY and not of O. One could then still derive lower and upper probabilities for this event, consistent with this data representation, by minimising and maximising, respectively, over all lower and upper probabilities for the same event based on the possible corresponding more detailed data representations which are in agreement with D_a, which in this simple case would be $(R : 1; O : 2)$ and $(Y : 1; O : 2)$.

Walley [16] ended his paper with the challenge to other researchers to develop and apply other methods, and to report their numerical answers. With this paper, we have responded to this challenge, and we share Walley's wish that this will lead to a wider discussion of this interesting and important inferential problem.

Appendix

For a closer investigation of the claim, in Section 2.1, that these NPI-based inferences lead to F-probability, we consider the two steps involved in the construction of our lower and upper probabilities separately. In Part a we look at a single configuration of equally spaced lines on the probability wheel which represent the observations, under the model assumption that each observation category is presented by only a single segment of the probability wheel so that lines representing observations in one category are 'next to each other', and in Part b we discuss the combination of all the configurations which are in agreement with our available observations under a given data representation, and with the assumption $\mathcal{A}_{(n)}$.

Part a: Let σ be a single configuration, which consists of a certain ordering of n lines representing k different categories on the probability wheel, under the condition that all the lines representing observations in the same category lay in one segment. Actually this means that we place k blocks of different colours on the probability wheel, where block r contains n_r lines of the corresponding category, for $r = 1, \ldots, k$. Applying $\mathcal{A}_{(n)}$ to one such a configuration, we obtain lower and upper limits $\underline{P}^{(\sigma)}(\cdot)$ and $\overline{P}^{(\sigma)}(\cdot)$, for the predictive probability of interest, which are totally monotone and totally alternating, and which produce an F-probability field. To prove this, note that the proofs of Theorem 1 and Theorem 2 in [2], where $A_{(n)}$ is considered on the real line, do not use the ordering of the intervals created by the data on the real line, and therefore these results can be adopted directly for the probability wheel representation for NPI-based inference based on multinomial data, and the assumption $\mathcal{A}_{(n)}$, as used in this paper. Note furthermore, that the fact that the resulting lower and upper probabilities are interval limits of an F-probability field, implies coherence and avoiding sure loss in Walley's sense [16].

Part *b*: We now consider all such configurations which are in agreement with the observations under a given data representation, and with $\mathcal{A}_{(n)}$. According to one of Walley's lower envelope theorems [15, Thm 2.6.3], by passing over to the minimal and maximal predictive interval probabilities over all such configurations, the properties of coherence and avoiding sure loss are preserved. A similar result is true for F-probability: With Σ as the set of all such configurations, $\underline{P}(\cdot) := \min_{\sigma\in\Sigma} \underline{P}^{(\sigma)}(\cdot)$ and $\overline{P}(\cdot) := \max_{\sigma\in\Sigma} \overline{P}^{(\sigma)}(\cdot)$ again constitute an F-probability field, conform to Weichselberger's concept of the union of F-probability fields [19, Lemma 2.7.12], see [1, Thm 3.2] for a proof in the context of total-monotonicity and belief functions.[3]

References

[1] T. Augustin. Generalized basic probability assignments. *International Journal of General Systems*, conditionally accepted.

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

[4] F.P.A. Coolen. Contribution to discussion of Walley [16]. *Journal of the Royal Statistical Society B*, 58: 43, 1996.

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

[6] F.P.A. Coolen and T. Augustin. Nonparametric predictive inference for multinomial data. *In preparation.*

[7] F.P.A. Coolen and K.J. Yan. Nonparametric predictive comparison of two groups of lifetime data. In: *ISIPTA'03 - Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, Bernard, Seidenfeld and Zaffalon (eds), Proceedings in Informatics 18, Carlton Scientific, 148-161, 2003.

[8] 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.

[9] P. Coolen-Schrijner and F.P.A. Coolen. Adaptive age replacement based on nonparametric predictive inference. *Journal of the Operational Research Society*, 55: 1281-1297, 2004.

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

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

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

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

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

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

[16] 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.

[17] P. Walley and J.-M. Bernard. Imprecise probabilistic prediction for categorical data. *Technical Report CAF-9901*, Laboratoire Cognition et Activités Finalisées, Université Paris 8, Saint-Denis, France, 1999.

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

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

[20] K. Weichselberger. The logical concept of probability and statistical inference. Conditionally accepted for *ISIPTA 2005.*

[3] The set of totally monotone set-functions is not closed under this process, so the $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$, presented in this paper, may be suspected not to be totally monotone, and even not two-monotone. This will be studied in more detail in [6], but has no consequence with respect to interpretation. However, it could lead to some numerical effort required for calculating corresponding conditional lower and upper probabilities.