Coolen, F. P. A.; Augustin, Thomas

Working Paper

A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories

Discussion Paper, No. 489

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Coolen, F. P. A.; Augustin, Thomas (2006) : A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories, Discussion Paper, No. 489, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1857

A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories

F.P.A. Coolen [a] T. Augustin [b]

[a] *Durham University, Department of Mathematical Sciences, Science Laboratories, Durham, DH1 3LE, UK*

[b] *Ludwig-Maximilians University, Department of Statistics, Ludwigstr. 33, D-80539 Munich, Germany*

Abstract

Nonparametric Predictive Inference (NPI) is a general methodology to learn from data in the absence of prior knowledge and without adding unjustified assumptions. This paper develops NPI for multinomial data where the total number of possible categories for the data is known. We present the general upper and lower probabilities and several of their properties. We also comment on differences between this NPI approach and corresponding inferences based on Walley's Imprecise Dirichlet Model.

Key words: Imprecise Dirichlet Model, imprecise probabilities, interval probability, known number of categories, lower and upper probabilities, multinomial data, nonparametric predictive inference, probability wheel.

1 Introduction

Many statistical data, e.g. from medicine to social and economics statistics, are multinomial, i.e. the observations fall into one of several unordered categories. In this paper we present a powerful approach to learning from multinomial data with at most $K \geq 3$ different categories, where K is explicitly known. The approach provides us with upper and lower probabilities for (a) future observation(s), and it appears to be an attractive alternative to Walley's Imprecise Dirichlet model [27], which has attracted considerable attention in a

Email addresses: frank.coolen@durham.ac.uk (F.P.A. Coolen), thomas@stat.uni-muenchen.de (T. Augustin).

variety of different applications (see, in particular, the survey by Bernard [4] and this special issue of International Journal of Approximate Reasoning.)

Our approach relies on the general framework of 'Nonparametric Predictive Inference' (NPI) [3,8], which is based on Hill's assumption $A_{(n)}$ [21]. By using the same variation of this assumption as presented in [9], called 'circular-$A_{(n)}$', our inference is closely related to our approach sketched in [9] where we explicitly do not assume any knowledge about the number of possible categories, apart from the information in the available data. A detailed and extensive presentation of NPI for multinomial data, considering all relevant aspects and containing detailed proofs and discussions of principles of general interval probabilistic statistical inference, is in preparation [10]. In the current paper, we present related results for the practically important case of a known number of possible categories, which is closer in nature to the traditional use of multinomial distributions. In comparison to the results without such knowledge [9], the inferences in this paper are either the same, or have less imprecision, in the latter case the lower and upper probabilities will be nested in the logical manner.

In Section 2, we give brief introductions to $A_{(n)}$, circular-$A_{(n)}$, interval probability and NPI, and to the model underlying our inferences [9,10]. The main results, NPI-based lower and upper probabilities for the next observation on the basis of multinomial data with a known number of possible categories, are presented in Section 3, where we also formulate some general properties of these inferences. In Section 4 these results are compared to the IDM and numerical examples are used to illustrate particular features of these inferences. In Section 5 some additional issues are discussed. An explanation of the derivation of the lower and upper probabilities is provided in an Appendix.

2 Nonparametric Predictive Inference and the underlying model

Hill [21] 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 [17], which provides direct posterior predictive probabilities [18]. Hill [21,22] 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 [3] prove that Nonparametric Predictive Inference (NPI) based on $A_{(n)}$ has strong consistency properties in the theory of interval probability [26,29,30]. In Theorem 1 it will be shown that the predictive lower and upper probabilities presented in this paper are internally consistent in the same, very strong sense.

In our model, we represent multinomial data as observations on a probability wheel, and hence as circular data. For such 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 *circular*-$A_{(n)}$, denoted by $\mathcal{A}_{(n)}$ [8,9]: 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 (classical) probability, so

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

Notice that neither the units of the circular data, nor the chosen 0-point on the circle, are relevant here. Clearly, $\mathcal{A}_{(n)}$ 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. Hence, NPI based on $\mathcal{A}_{(n)}$ has the same consistency properties as shown in [3] for such inference based on $A_{(n)}$.

In this paper, as in [9,10], we use $\mathcal{A}_{(n)}$ combined with an assumed underlying representation of multinomial data as outcomes of spinning a probability wheel. As we wish not to make further assumptions about the probability mass $1/n$ per interval I_j, our predictive inferences are again in the form of interval probabilities [3,26,29,30], 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 $\mathcal{A}_{(n)}$ and in accordance with the probability wheel model, according to De Finetti's fundamental theorem of probability [17], and the upper probability is the minimum upper bound consistent in this way. From a subjective point of view as advocated by Walley [26], 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 Section 3 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 and due to the fact that they bound the observed relative frequencies (Theorem 1), but are also consistent if more observations are added to the data. Such considera-

tions will be discussed in detail in [10], together with the underlying model and the principles leading to, and detailed justification of, lower and upper probabilities presented in Section 3 and in [9]. 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 (3) and (4) 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 [19] 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, and the fact that there are at most K possible categories, hence K different segments on the probability wheel. 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 adds to imprecision of our inferences, in the sense that our lower and upper probabilities are optimal bounds over all configurations of these K possible segments on the probability wheel.

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 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 $\mathcal{A}_{(n)}$, 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 segments 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 Lower and upper probabilities

In a standard multinomial setting, observations belong to categories, with no natural relationships or orderings between these categories. We assume that there is a known number of possible categories, denoted by K, and 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 [7], which leads to slightly less imprecision than a representation on a circle as in this paper. We assume that each observation can be assigned to a category with certainty. Our inferences in this paper are based on the assumption that $n \geq 1$ observations are available, and the inferences are predictive, focussing on a single future observation denoted by Y_{n+1}, which is assumed to be exchangeable with the n observations so far.

We denote the $K \geq 3$ possible categories by $C_1, \ldots, C_K$. 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 event of interest in this paper can generally be denoted by

$$Y_{n+1} \in \bigcup_{j \in J} C_j \tag{2}$$

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 NPI-based lower and upper probabilities for event (2), based on the n observations [1], the assumption $\mathcal{A}_{(n)}$ and our probability wheel model, are

$$\underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n} \left(\sum_{j \in OJ} n_j - r + \max(2r + l - K, 0) \right) \tag{3}$$

and

$$\overline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n} \left(\sum_{j \in OJ} n_j - r + \min(2r + l, k) \right) \tag{4}$$

For the two trivial events, the NPI-based lower and upper probabilities are obvious. If $J = \{1, \ldots, K\}$, the upper probability of event (2) is equal to 1, in line with (4), and also the lower probability (3) is trivially defined as 1, which is fully in line with the probability wheel model which underlies our inferences. Similarly, if $J = \emptyset$, the lower probability of event (2) is equal to 0, in line with (3), and the upper probability (4) is defined as 0. In our discussion below, we will not explicitly mention these trivial events anymore.

To derive these lower and upper probabilities we consider all possible configurations σ on the probability wheel, apply $\mathcal{A}_{(n)}$ 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. In the Appendix the lower and upper predictive probabilities (3) and (4) are directly derived by constructing those configurations that minimize $\underline{P}_\sigma(\cdot)$ and maximize $\overline{P}_\sigma(\cdot)$.

We now directly turn to some fundamental properties of our inferences:

Theorem 1:

The lower and upper probabilities (3) and (4) satisfy the following properties:

i) (Conjugacy) For all $J \subseteq \{1, \ldots K\}$:

$$\underline{P} \left(Y_{n+1} \in \bigcup_{j \in J} C_j \right) = 1 - \overline{P} \left(Y_{n+1} \in \bigcup_{j \in \{1, \ldots, K\} \setminus J} C_j \right) .$$

[1] All probabilities considered here are predictive given the first n observations. So we do not explicitly mention the dependence on the first n observations in the notation.

ii) For all $J \subseteq \{1, \ldots, K\}$:

$$\underline{P}\left(Y_{n+1} \in \bigcup_{j \in J} C_j\right) \leq \frac{\sum_{j \in J} n_j}{n} \leq \overline{P}\left(Y_{n+1} \in \bigcup_{j \in J} C_j\right).$$

iii) If n varies, and $\underline{P}^{(n)}(\cdot)$ and $\overline{P}^{(n)}(\cdot)$ are the corresponding lower and upper probabilities based on n observations, then, for every $J \subseteq \{1, \ldots, K\}$,

$$\lim_{n \to \infty} \underline{P}^{(n)}\left(Y_{n+1} \in \bigcup_{j \in J} C_j\right) = \lim_{n \to \infty} \overline{P}^{(n)}\left(Y_{n+1} \in \bigcup_{j \in J} C_j\right).$$

iv) $P(\cdot) = [\underline{P}(\cdot), \overline{P}(\cdot)]$ is F-probability in the sense of Weichselberger (2001), i.e. with $p(\cdot)$ denoting classical probabilities and

$$\mathcal{M} := \Big\{ p(\cdot) \Big| \underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) \leq p(Y_{n+1} \in \bigcup_{j \in J} C_j) \leq \overline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j),\ \forall J \subseteq \{1, \ldots, K\} \Big\} \tag{5}$$

as the so-called structure consisting of all classical probabilities being in accordance with $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$, one obtains

$$\underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \min_{p(\cdot) \in \mathcal{M}} p(Y_{n+1} \in \bigcup_{j \in J} C_j)$$

and

$$\overline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \max_{p(\cdot) \in \mathcal{M}} p(Y_{n+1} \in \bigcup_{j \in J} C_j).$$

v) $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$ are coherent lower and upper probabilities in the sense of Walley (1991).

Sketch of proof: i) to iii) can be demonstrated by elementary methods. For iv) it is helpful to realize that $\mathcal{A}_{(n)}$ based on a certain arbitrary configuration σ can be formally described by a basic probability assignment $m_\sigma(\cdot)$, leading to a belief function $\underline{P}_\sigma(\cdot)$ and a plausibility function $\overline{P}_\sigma(\cdot)$ with structure $\mathcal{M}_\sigma$. Taking the lower and upper envelope over all possible configurations leads to F-probability [2, Th. 2.3.2] [2], which is here, by the lower envelope theorem [26, p.134], equivalent to coherence. ■

The theorem formulates important properties of internal consistency and rationality of our model. Property ii) distinguishes our model as a proper generalization of the naïve predictive learning where simply relative frequencies are assigned: Our model contains the relative frequencies but also reflects, by

[2] This way to derive $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$ shows that our inferences are a special case of generalized basic probability assignments ([2]), see also [16] for a related concept.

the imprecision, the amount of information on which the inferences are based. This imprecision vanishes if, and only if, the sample size tends to infinity (cf. Property iii), so that with full information we eventually learn the true proportions of the segments on the probability wheel.

Property i) shows that upper and lower probability fit to each other in a complementary way. Far beyond this minimal requirement our inference leads to F-probability in the interval probability theory of Weichselberger [3,29,30]. 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 the induced set of classical ('precise') probabilities, and so the bounds make use of the available information in a perfect manner; they are neither too wide nor do they add unjustified additional assumptions to our inferences. Additionally, since on finite spaces the F-probability property coincides with coherence in Walley's sense [26], our bounds are also perfectly rational from the behavioral point of view.

If we want to apply our model for predictive decision making or classification, for instance, we have to go a step further and associate real-valued outcomes with every category $C_1, \ldots, C_K$, i.e. we want to consider random quantities $X : \{C_1, \ldots, C_K\} \longrightarrow \mathbb{R}$. To determine their lower and upper expectation there are two different ways to proceed:

a) The *direct method* copies the derivation of $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$ from (3) and (4) by replacing probability with expectation. So, we consider firstly every configuration σ on the probability wheel separately, calculate the corresponding lower and upper expectation $\underline{\mathbb{E}}_\sigma X$ and $\overline{\mathbb{E}}_\sigma X$ and then consider the envelope over all configurations $\sigma \in \Sigma$, resulting in

$$\underline{\mathbb{E}}\, X := \min_{\sigma \in \Sigma} \underline{\mathbb{E}}_\sigma X \quad \text{and} \quad \overline{\mathbb{E}}\, X := \max_{\sigma \in \Sigma} \overline{\mathbb{E}}_\sigma X\,. \tag{6}$$

b) The *indirect method* uses the predictive lower and upper probabilities from (3) and (4) as the fundamental building blocks, from which then the lower and upper expectations are derived. As a consequence, we first determine our lower and upper probabilities $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$ in accordance with the model, and then use the corresponding structure $\mathcal{M}$ (cf. (5)) to define lower and upper expectations:

$$\underline{\mathbb{E}}_{\mathcal{M}} X := \min_{p \in \mathcal{M}} \mathbb{E}_p X \quad \text{and} \quad \overline{\mathbb{E}}_{\mathcal{M}} X := \max_{p \in \mathcal{M}} \mathbb{E}_p X\,. \tag{7}$$

From general theory (e.g., [26, p.81]) it is well known that in general

$$\underline{\mathbb{E}}_{\mathcal{M}} X \leq \underline{\mathbb{E}}\, X \quad \text{and} \quad \overline{\mathbb{E}}\, X \leq \overline{\mathbb{E}}_{\mathcal{M}} X \tag{8}$$

with strict inequalities being possible, i.e., from the viewpoint of the direct method, the indirect method could lead to substantial loss of information. It

is therefore quite a strong internal consistency property (closedness property) of our model that the inner and the outer methods coincide. Moreover, quite important from the applied point of view, we can give convenient expressions to calculate the lower and the upper expectations as simply weighted sums instead of solutions to linear optimization problems:

Theorem 2:

Consider a random quantity $X : \{C_1, \ldots, C_K\} \to \mathbb{R}$. Then, with the notation from (6) and (7)

i)

$$\underline{\mathbb{E}}\, X = \underline{\mathbb{E}}_{\mathcal{M}} X = \sum_{J \subseteq \{1,\ldots,K\}} m\left(Y_{n+1} \in \bigcup_{j \in J} C_j\right) \min_{j \in J} X(j)$$

and

$$\overline{\mathbb{E}}\, X = \overline{\mathbb{E}}_{\mathcal{M}} X = \sum_{J \subseteq \{1,\ldots,K\}} m\left(Y_{n+1} \in \bigcup_{j \in J} C_j\right) \max_{j \in J} X(j),$$

where $m(\cdot)$ is the Moebius inverse of $\underline{P}(\cdot)$, i.e. for every $J \subseteq \{1, \ldots, K\}$,

$$m\left(Y_{n+1} \in \bigcup_{j \in J} C_j\right) = \sum_{I \subseteq J} (-1)^{|J \setminus I|} \underline{P}\left(Y_{n+1} \in \bigcup_{i \in I} C_i\right). \tag{9}$$

ii) Let furthermore $x_{(1)} < x_{(2)} < \ldots < x_{(q)}$, $q \leq K$, be the distinct values of the image of X (ordered in increasing magnitude), and define $J(i) = \left\{ j \in \{1, \ldots, K\} \middle| X(j) = x_{(i)} \right\}$ and $I(i) := \cup_{t=1}^{i} J(t)$, $i = 1, \ldots, q$, then

$$\underline{\mathbb{E}}\, X = \underline{\mathbb{E}}_{\mathcal{M}} X = x_{(1)} + \sum_{i=2}^{q} \left(x_{(i)} - x_{(i-1)}\right) \cdot \underline{P}\left(Y_{n+1} \in \bigcup_{s \in I(i)} C_s\right)$$

and

$$\overline{\mathbb{E}}\, X = \overline{\mathbb{E}}_{\mathcal{M}} X = x_{(1)} + \sum_{i=2}^{q} \left(x_{(i)} - x_{(i-1)}\right) \cdot \overline{P}\left(Y_{n+1} \in \bigcup_{s \in I(i)} C_s\right).$$

Sketch of proof: For the technical handling of the argument to be given here, it is helpful to identify every configuration σ on the probability wheel with a permutation of $\{1, \ldots, K\}$, where (some of) the not yet seen colours, i.e. some indices, may simply not be visible, so they can be interpreted as corresponding to segments of size 0 on the probability wheel. Denote the set of all permutations again by Σ. Then the proof relies on the following lemma:

Lemma 1:
For every increasing sequence $\mathfrak{S}$ of index sets $J(i) \subseteq \{1, \ldots, K\}$, $J(i) \subseteq J(j)$,

$i \leq j$, there is a permutation $\sigma_0 \in \Sigma$ such that

$$\overline{P}\left(Y_{n+1} \in \bigcup_{j \in J(i)} C_j\right) = \overline{P}_{\sigma_0}\left(Y_{n+1} \in \bigcup_{j \in J(i)} C_j\right) \text{ for all } i\,.$$

Sketch of proof of Lemma 1: Note that σ_0 is required to be the same for all elements of the sequence, and so the lemma says that for every sequence there is a 'favorable configuration' in which for *all* elements of the sequence the upper probability is attained *simultaneously*. Such a sequence can indeed be constructed, namely by separating neighboring elements of the sequence as long as possible: If, without loss of generality,

$$J(1) = \{1\}, \quad J(2) = \{1,2\}, \quad \ldots, \quad J(K) = \{1,\ldots,K\}\,,$$

then the appropriate permutation σ_0 is obtained by the following ordering of the colours on the circle:

$$\{1\}, \quad \{K\}, \quad \{2\}, \quad \{K-1\}, \quad \{3\}, \quad \{K-2\}, \quad \text{etc.},$$

because then for every element of the sequence $\mathfrak{S}$ the upper probability, as long as it is smaller than 1, increases for every $j \in J(i)$ by $n_j + 1$ if colour j is among the categories seen so far and by 1 if not. This exactly coincides with $\overline{P}(\cdot)$ for all events corresponding to elements of $\mathfrak{S}$. ■

An immediate consequence of the lemma is that $\underline{P}(\cdot)$ is two-monotone and $\overline{P}(\cdot)$ is two-alternating. Therefore, by applying results on the Choquet integral (e.g. [5]), the lower and upper expectations $\underline{\mathbb{E}}_{\mathcal{M}} X$ and $\overline{\mathbb{E}}_{\mathcal{M}} X$ can be determined as described.

To show the equality on the left hand side, note firstly that, for every configuration σ, $\underline{\mathbb{E}}_\sigma X$ and $\overline{\mathbb{E}}_\sigma X$ arise from optimizing $\mathbb{E}_p X$ over the corresponding structure $\mathcal{M}_\sigma$ (see the sketch of proof of Theorem 1), and so

$$\underline{\mathbb{E}}\, X = \min_{\sigma \in \sigma} \underline{\mathbb{E}}_\sigma X = \min_{\sigma \in \Sigma} \min_{p \in \mathcal{M}_\sigma} \mathbb{E}_p X\,.$$

Secondly, the vertices (extremal points) of the structure $\mathcal{M}$ of an F-probability with two-monotone lower interval limit $\underline{P}(\cdot)$ are obtained by considering all sequences of the form described in Lemma 1 (cf., e.g., [5]). Consequently, Lemma 1 tells us that $\mathcal{E}(\mathcal{M}) \subseteq \bigcup_{\sigma \in \Sigma} \mathcal{M}_\sigma$, which leads to

$$\underline{E}_{\mathcal{M}} X = \min_{p \in \mathcal{E}(\mathcal{M})} \mathbb{E}_p X \geq \min_{\sigma} \min_{p \in \mathcal{M}_\sigma} \mathbb{E}_p X\,.$$

This gives, together with (8), the equivalence of the direct and the indirect methods. ■

4 Comparison with IDM and examples

In this section we compare our NPI-based inferences for multinomial data with K possible categories to Walley's Imprecise Dirichlet Model [27], in particular by focussing on some specific situations. We illustrate our method, also to appreciate the differences to the IDM [3], via some numerical examples. The lower and upper probabilities for the general event (2), according to Walley's IDM [27], based on the n observations as described above, are

$$\underline{P}_{IDM}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n+s} \left(\sum_{j \in OJ} n_j \right) \tag{10}$$

for $J \neq \{1, \ldots, K\}$, and

$$\overline{P}_{IDM}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n+s} \left(\sum_{j \in OJ} n_j + s \right) \tag{11}$$

for $J \neq \emptyset$, with s a positive constant to be chosen independently of the data. Walley [27] advocates the use of small values such as $s = 1$ or $s = 2$ to align the resulting inferences with several frequently suggested Bayesian and classical inferences. Walley [27] states as important advantage of the IDM 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. [4] For non-singletons, our lower and upper probabilities (3) and (4) do not generally satisfy the RIP, but as will become obvious from the examples below we do not see this as a disadvantage of our model.

The discussants to Walley's paper [27] 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 $\frac{1}{1+s}$. The suggested small values of s, in particular $s = 1$ or $s = 2$, lead to intuitively surprisingly high values for this lower probability. As discussed below, in our NPI-based approach, this lower probability is 0. (2) (RIP:) The IDM predictive lower and upper probabilities depend only on the observed frequency of that category and the total number of observations. This is not the case for our NPI-based

[3] The Imprecise Dirichlet-Multinomial Model [28] gives the same lower and upper probabilities (10) and (11) for the general event (2) as the IDM, so it does not provide an alternative solution with regard to the issues mentioned in this paper on the comparison of our NPI-based method and the IDM.

[4] Far beyond this the IDM could in some sense be characterized as the only model satisfying the RIP [15].

lower and upper probabilities, we illustrate this explicitly in the examples below. (3) The IDM upper probabilities for events that the next observation is in an as yet unseen category do not depend on the number of categories seen so far. This is not the case for our NPI-based upper probability, as illustrated in Example 2 below. A further important advantage of our NPI-based approach over the IDM approach appears if one does not know the total number of possible categories, and wishes to distinguish in the event of interest between fully defined categories that have not yet been observed, and any new category occurring at the next observation. Our NPI-based lower and upper probabilities for this situation are presented in [9], where the corresponding inferences are also compared in detail with the IDM.

We illustrate our lower and upper probabilities (3) and (4) for some special cases of the general event (2) and available data, also commenting on the corresponding IDM lower and upper probabilities where useful to highlight differences and similarities.

It is clearly of interest to consider the NPI-based lower and upper probabilities for events containing only a single category. Let us begin with the case that this one category has already been observed, so $r = 1,\ l = 0$, and without loss of generality let us assume that the category of interest is C_1, so $n_1 \geq 1$, then

$$\underline{P}(Y_{n+1} \in C_1) = \frac{n_1 - 1}{n} \tag{12}$$

as we assumed throughout that $K \geq 3$ (the same would hold in NPI based on this probability wheel model for $K = 2$), and

$$\overline{P}(Y_{n+1} \in C_1) = \min\left(\frac{n_1 + 1}{n}, 1\right) \tag{13}$$

again as we assumed that $K \geq 3$ (with the same added comment as above). The IDM lower and upper probabilities for this event are $n_1/(n + s)$ and $(n_1+s)/(n+s)$, respectively. Note that our lower probability (12) only becomes positive if $n_1 > 1$, so in this case the NPI-based inference is quite conservative when compared to the IDM [6].

Secondly, if this one category of interest has not yet been observed, so $r = 0,\ l = 1$, and without loss of generality let us assume that the category of interest is C_K, then

$$\underline{P}(Y_{n+1} \in C_K) = 0 \tag{14}$$

and

$$\overline{P}(Y_{n+1} \in C_K) = \frac{1}{n} \tag{15}$$

For this event, the IDM lower and upper probabilities are 0 and $s/(n + s)$, respectively. We see from (12)-(15) that the RIP actually holds in our approach for events involving only a single category.

If all r observed categories in the event of interest have been observed exactly once, so $n_j = 1$ for all $j \in OJ$, then

$$\underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n} \max(2r + l - K, 0) \tag{16}$$

which exceeds 0 if and only if $K - r - l < r$. In this latter case, there are not enough categories available which are not in the event of interest and for which the corresponding segments in the probability wheel model can be used to separate all r segments corresponding to observed categories in the event of interest (see the explanation of the derivation of (3) and (4) in the Appendix). We illustrate and discuss this lower probability in Example 1. The IDM lower probability for this event is $r/(n+s)$, which for small values of s is larger than the NPI-based lower probability (16).

For events containing only categories which have not yet been observed, so $r = 0$ and $1 \leq l \leq K - k$, the upper probability is

$$\overline{P}(Y_{n+1} \in \bigcup_{j \in UJ} C_j) = \frac{1}{n} \min(l, k) \tag{17}$$

while the corresponding NPI-based lower probability is 0, and the IDM upper probability for this event is $s/(n + s)$ for all values of l and k. We illustrate and discuss this upper probability in Example 2.

Example 1.
Suppose that $K = 10$, $n = 20$, and $k = 5$, with $n_1 = 16$ observations in C_1, and $n_j = 1$ for $j = 2, \ldots, 5$ in each of categories C_2 to C_5. Suppose that interest is in the events $Y_{21} \in \bigcup_{j \in \{2,\ldots,5+l\}} C_j$ for $l \in \{0, \ldots, 5\}$, so that the next observation belongs to any of the categories with a single observation so far or to any of the first l not yet observed categories. By (16) the lower probability for this event is 0 for $l \leq 2$. However, for $l \in \{3, 4, 5\}$ this lower probability is positive, namely $(l-2)/20$. Of course, for $l = 5$ the event of interest is just the complementary event to $Y_{n+1} \in C_1$, and this lower probability of $3/20$ then also follows by the conjugacy property from the fact that $\overline{P}(Y_{n+1} \in C_1) = 17/20$ using (13).

So, in most cases the lower probability for the next observation to belong to categories which have been observed at most once is 0, but if many such categories are included in the event of interest, in comparison to the number of included categories for which the data contain more than one observation, then this lower probability can become positive in our NPI-based approach.

The IDM gives $\underline{P}_{IDM}(Y_{21} \in \bigcup_{j \in \{2,\ldots,5+l\}} C_j) = \frac{4}{20+s}$ for all $l \in \{0, \ldots, 5\}$, which corresponds naturally to the IDM upper probability of $(16 + s)/(20 + s)$ for the event $Y_{n+1} \in C_1$, yet it may be deemed somewhat surprising that it does

not at all depend on the value of l. Most noticeably, of course, is the equality of these IDM lower probabilities for the cases $l = 0$ and $l = 5$, so it does not matter to the IDM whether none or all not yet observed categories are included in the event of interest.

In this example, one could consider the NPI-based approach to be a bit more conservative for these events than the IDM approach, as long as small values of s are used for the latter.

Example 2.
Suppose that interest is in events expressing that Y_{n+1} does not belong to any of the categories of the first n observations. The NPI-based lower probability for any such an event is 0, the upper probability is given by (17). The most general formulation of this event is $Y_{n+1} \in \bigcup_{j\in\{k+1,\ldots,K\}} C_j$, but instead one may have an explicit interest in a subset of the not yet observed categories, $Y_{n+1} \in \bigcup_{j\in UC} C_j$, with l as before the number of elements of UC. Suppose that $K = 40$ different categories are possible, and $n = 200$ observations are available which belong to $k = 5$ different categories, C_1 to C_5. The corresponding NPI-based upper probability that Y_{201} belongs to any not yet observed category is $5/200$, while for the event that it belongs to any specific such category, so any from C_6 to C_{40}, the upper probability is $1/200$, and for any pair of these 35 not yet observed categories the corresponding upper probability is $2/200$, and so on up to the upper probability $5/200$ for any subset containing 5 or more of these unobserved categories.

If, instead, the first $n = 200$ observations had been in $k = 20$ different categories, C_1 to C_{20}, then the corresponding NPI-based upper probability that Y_{201} belongs to any not yet observed category is $20/200$, while the similar case with $k = 25$ categories already observed would lead to the value $15/200$ for this upper probability. For subsets consisting of l of the unobserved categories, the corresponding upper probabilities are $l/200$ for $l \leq k$, and $k/200$ for $l \geq k$. These values reflect both how many different categories have already been observed, and how many unobserved categories are still available. The maximum possible value of such an upper probability is attained for the case where the number k of observed categories is equal to the number of unobserved categories.

The IDM gives upper probability $s/(200+s)$ for all the events in this example, as it does not distinguish between such events with different subsets of the unobserved categories. It also does not take into account k, the number of observed categories in the 200 observations so far.

Example 3.
Table 1 presents the NPI and IDM (with $s = 1$) lower and upper probabilities

for all non-trivial events of interest on Y_{11}, in the case with $K = 4$ categories and $n = 10$ observations, with all categories observed and $n_j = j$ observations in category C_j, for $j = 1, \ldots, 4$.

J	$\underline{P}$	$\overline{P}$	$\underline{P}_{IDM}$	$\overline{P}_{IDM}$
1	0/10	2/10	1/11	2/11
2	1/10	3/10	2/11	3/11
3	2/10	4/10	3/11	4/11
4	3/10	5/10	4/11	5/11
1,2	1/10	5/10	3/11	4/11
1,3	2/10	6/10	4/11	5/11
1,4	3/10	7/10	5/11	6/11
2,3	3/10	7/10	5/11	6/11
2,4	4/10	8/10	6/11	7/11
3,4	5/10	9/10	7/11	8/11
1,2,3	5/10	7/10	6/11	7/11
1,2,4	6/10	8/10	7/11	8/11
1,3,4	7/10	9/10	8/11	9/11
2,3,4	8/10	10/10	9/11	10/11

Table 1. NPI and IDM lower and upper probabilities (Example 3)

This basic example highlights that for the IDM, imprecision, which is the difference between corresponding upper and lower probabilities, does not depend on the event of interest [5] (we again do not consider the trivial events in this discussion), whereas imprecision does depend on the event in the NPI-based approach, with imprecision in this example larger in case J has two elements than for one element (or three of course, by conjugacy which is also illustrated throughout in Table 1). In most situations in our approach, imprecision is larger for events involving unions of categories than for events involving single categories. In this example, the values are pretty similar, with the NPI-based approach a bit more conservative due to the small chosen value $s = 1$ in the IDM. In most situations with observations available in all categories, this will be the case, as the most noticeable differences between the NPI and IDM approaches occur in situations as discussed in Examples 1 and 2. Both methods also have the intuitively logical properties that the lower and upper prob-

[5] For this reason a vivid non-Bayesian interpretation of the IDM sees it as contaminated relative frequencies [23].

abilities always bound the corresponding relative frequency of the event of interest in the data, and that the lower and upper probabilities converge to this relative frequency if the numbers of observations in the categories become large. Moreover, the equivalence of the direct and indirect method in assigning expectations (cf. Theorem 2) can also be shown to hold for the IDM [25].

5 Discussion

In recent years, Walley's IDM [27] has received increasing attention and gained popularity for a variety of applications, as is clear from [4] and this special issue of International Journal of Approximate Reasoning. Although our NPI-based inferences, as presented here and in [9,10], are close to corresponding IDM results in situations with lots of data and known categories, they can differ substantially in other situations as highlighted in the examples in this paper and in [9]. It is an interesting topic for future research to develop applications based on our NPI-based method, for example classification, and to compare their results with corresponding outcomes from the IDM (e.g.[1,24,31]). This is also necessary to investigate the practical relevance of our proposed method.

Although our NPI approach is naturally presented in terms of a single future observation, it is conceptually straightforward to extend it to any number of future observations, via sequential arguments. Any statistical approach must have consistency properties for updating and conditioning. We discuss these important features in detail in [3] and [10], where it is emphasized that these are very different actions. Updating involves learning from more observations, and adapting inferences to this. This is naturally done in the NPI framework by taking all new data into account together with the previously available data, and basing predictive inference on the appropriate $A_{(\tilde{n})}$ or $\textcircled{A}_{(\tilde{n})}$ assumption with $\tilde{n}$ the new total number of observations. Conditioning, on the other hand, typically involves taking specific additional information on the random quantity of interest, Y_{n+1}, into account. For both these actions, strong consistency results hold for NPI, more details will be presented in [10].

As discussed in Section 4, and in more detail in [9], the RIP does not generally hold for the NPI-based inferences. Hence, our inferences can depend on the choice of categories used to represent the data. We believe that this is a natural feature of statistical inference based on lower and upper probabilities. 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, and 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 possibly more, but never less, predictive imprecision in case of a more detailed data representation. Generally speaking, our NPI-based inferences for multinomial data are minimally imprecise if, for the event of interest, the data available are only recorded in a binary mode, so counting how often the event did or did not occur in the past. 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. In [10] more attention will be paid to this feature, suggesting a general property relating imprecision in inference to the level of detail of the data representation (sample space) that is weaker than the RIP, but also trivially satisfied by any method satifying the RIP.

The IDM has some important advantages over our NPI-based approach. In particular, as it is a parametric model in the Bayesian framework, it allows a 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 [11–14,20].

Acknowledgement:
The authors thank Gert de Cooman for interesting and stimulating discussions about predictive inference and interval probability. The second author is grateful to the Department of Mathematics at Durham University, and to the DFG within the frame of the Sonderforschungsbereich 386, for financial support when visiting the first author.

Appendix: Derivation of (3) and (4)

For the general event (2) considered, and notation used, in this paper, there are $r + l$ categories in the event of interest, and $K - r - l$ categories not in the event of interest. With regard to their representation via segments

on the probability wheel, as underlies our model and inferences, these latter categories play an important role with regard to specific configurations for which, combined with $\mathcal{A}_{(n)}$, the lower and upper probabilities (3) and (4) are obtained. The idea is straightforward, and identical in nature to that described in [9]. Detailed proofs, both for the case with known K and with unknown number of possible categories [9] will be given in [10].

We first consider the lower probability (3), and separate two cases. First, if $r \leq K - r - l$, so the number of possible categories not in the event of interest is not less than the number of already observed categories in that event, then there are sufficient segments possible on the probability wheel to enable configurations such that no segments corresponding to different already observed categories are next to each other. Of course, the segments falling in between lines representing the same category C_j, for $j \in OJ$ (there are no such segments for categories C_j with $j \in UC$), all represent predictive probabilities $1/n$ for Y_{n+1} that cannot be moved away from the event of interest, but no further segments must definitely belong to any categories in $\bigcup_{j \in J} C_j$. This clearly leads to

$$\underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \sum_{j \in OJ} \frac{n_j - 1}{n} = \frac{1}{n} \sum_{j \in OJ} n_j - r. \tag{18}$$

Secondly, if $r > K - r - l$ then not all previously observed C_j in the event of interest can be separated by categories not in the event of interest, the best we can do is to separate as many as possible, so use all $K - r - l$ categories not in the event of interest to separate as many of the r segments as possible. This leaves $r - (K - r - l) = 2r + l - K$ segments, each between two lines representing previous observations in different categories, that cannot be separated by categories not in the event of interest, and hence the probability masses $1/n$ assigned by $\mathcal{A}_{(n)}$ to each of these segments must also be taken into account, leading to

$$\underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{2r + l - K}{n} \tag{19}$$

Clearly, together with the lower probability for the first case above, this gives the general expression (3) for the lower probability.

Now we consider the upper probability (4). For this upper probability, all possible categories that have not yet been observed and that are not in the event of interest play no role, as effectively we can consider them either absent, or represented by a segment with area 0 on the circle. Hence, we only need to consider the probability mass that cannot be assigned to the event of interest, which is due to the $k - r$ already observed categories which are not in the event of interest. Again we consider two cases. First, if $k - r \leq r + l$, so if $k \leq 2r + l$,

then there are sufficient categories in the event of interest to ensure that no two observed categories not in it have to be next to each other. Hence, there are configurations for which all k segments in between two lines representing different observations on the probability wheel have at least one of these two lines belonging to a segment representing a category in the event of interest, and therefore the probability mass $1/n$ in each such segment can be assigned to the event of interest. This leads directly to

$$\overline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{k}{n}. \tag{20}$$

Secondly, if $k - r > r + l$, so if $k > 2r + l$, then some segments between two neighbouring lines representing observations in different categories cannot be included in the event of interest, as there are more such segments than there are categories (either observed or not) in the event of interest. Clearly, there are $k - r - (r + l) = k - 2r - l$ such segments that cannot be included in the event of interest, so from the k segments that are still free to be assigned after all segments uniquely assigned to a single category C_j have been assigned their probability masses $1/n$, we can assign the probability masses $1/n$ of $k - (k - 2r - l) = 2r + l$ such segments to the event of interest. Hence, in this case the upper probability is

$$\overline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{2r + l}{n}. \tag{21}$$

Clearly, together with the upper probability for the case above, this gives the general expression (3) for the upper probability.

References

[1] J. Abellán, S. Moral. Upper entropy of credal sets. Applications to credal classification. *Int. J. Approx. Reason.* 39 (2-3), 235-255, 2005

[2] T. Augustin. Generalized basic probability assignments. *Int. J. Gen. Syst.*, 34 (4): 451-463, 2005.

[3] T. Augustin and F.P.A. Coolen. Nonparametric predictive inference and interval probability. *J. Stat. Plan. Infer.*, 124 (2): 251-272, 2004.

[4] J.-M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *Int. J. Approx. Reason.*, 39 (2-3): 123-150, 2005.

[5] A. Chateauneuf, J.Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of Moebius inversion. *Math. Soc. Sci.*, 17 (3): 263-283, 1989.

[6] F.P.A. Coolen. Discussion of the paper by Walley [27], J. Roy. Stat. Soc. B 58 (1) (1996) 43.

[7] F.P.A. Coolen. Low structure imprecise predictive inference for Bayes' problem. *Stat. Probabil. Lett.*, 36 (4): 349-357, 1998.

[8] F.P.A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, to appear, 2006.

[9] F.P.A. Coolen and T. Augustin. Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. *ISIPTA'05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications*, Fabio G. Cozman, Robert Nau and Teddy Seidenfeld (Editors). (Published by SIPTA), 125-134, 2005.

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

[11] 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.

[12] F.P.A. Coolen and K.J. Yan. Nonparametric predictive inference with right-censored data. *J. Stat. Plan. Infer.*, 126 (1): 25-54, 2004.

[13] F.P.A. Coolen and P. Coolen-Schrijner. Nonparametric predictive comparison of proportions. *J. Stat. Plan. Infer.*, to appear.

[14] P. Coolen-Schrijner and F.P.A. Coolen. Adaptive age replacement based on nonparametric predictive inference. *J. Oper. Res. Soc.*, 55 (12): 1281-1297, 2004.

[15] G. de Cooman, On the representation invariance principle (working title), in preparation, 2006.

[16] G. de Cooman, E. Miranda, I. Couso. Lower previsions induced by multi-valued mappings. *J. Stat. Plan. Infer.*, 133 (1): 173-197, 2004.

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

[18] A.P. Dempster. On direct probabilities. *J. Roy. Stat. Soc. B*, 25: 100-110, 1963.

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

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

[21] B.M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *J. Am. Stat. Assoc.*, 63 (322): 677-691, 1968.

[22] B.M. Hill, B.M. 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, 211-241, 1988.

[23] T. Seidenfeld, L. Wassermann, Discussion of the paper by Walley [27], J. Roy. Stat. Soc. B 58 (1) (1996) 49.

[24] C. Strobl, Variable selection in classification trees based on imprecise probabilities. *ISIPTA'05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications*, Fabio G. Cozman, Robert Nau and Teddy Seidenfeld (Editors). (Published by SIPTA), 340-348, 2005.

[25] L.V. Utkin, T. Augustin. Decision making under incomplete data using the imprecise Dirichlet model. Conditionally accepted for *International Journal of Approximate Reasoning,* 2006.

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

[27] P. Walley. Inferences from multinomial data: learning about a bag of marbles (with discussion). *J. Roy. Stat. Soc. B*, 58 (1): 3-57, 1996.

[28] 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.

[29] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *Int. J. Approx. Reason.*, 24: 149-170, 2000.

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

[31] M. Zaffalon, Credible classification for environmental problems. *Environmental Modelling & Software,* 20(8): 1003-1012.