On nonparametric predictive inference and objective Bayesianism

F.P.A. Coolen
Department of Mathematical Sciences
University of Durham
Durham, DH1 3LE, UK
Frank.Coolen@durham.ac.uk

Abstract

This paper consists of three main parts. First, we give an introduction to Hill's assumption $A_{(n)}$ and to theory of interval probability, and an overview of recently developed theory and methods for nonparametric predictive inference (NPI), which is based on $A_{(n)}$ and uses interval probability to quantify uncertainty. Thereafter, we illustrate NPI by introducing a variation to the assumption $A_{(n)}$, suitable for inference based on circular data, with applications to several data sets from the literature. This includes attention to comparison of two groups of circular data, and to grouped data. We briefly discuss such inference for multiple future observations. We end the paper with a discussion of NPI and objective Bayesianism.

Keywords: $A_{(n)}$; circular data; exchangeability; grouped data; imprecise probabilities; interval probability; objective Bayesianism.

1 Introduction

Hill [44] proposed the assumption $A_{(n)}$ to provide direct probabilities for future observations [30, 35], based on data consisting of n observations on the real-line. In later work [45, 46], Hill discussed $A_{(n)}$ as a basis for Bayesian nonparametric predictive inference, and proved that $A_{(n)}$ fits into the general framework of Bayesian statistics by using a, rather complicated, splitting process to provide the prior distribution. We give a brief overview of $A_{(n)}$ in Section 2 of this paper.

This assumption $A_{(n)}$ is not generally sufficient to derive precise predictive probabilities for events of interest. However, theory of interval probability [61, 62], also known as 'imprecise

probabilities' [57], generalizes theory of precise (or 'classical') probability via the use of lower and upper probabilities, and provides a suitable framework for $A_{(n)}$-based nonparametric predictive inference (NPI) [3]. We give a brief overview of interval probability in Section 3, and of NPI in Section 4.

NPI uses the close relation between $A_{(n)}$ and finite exchangeability, and can be based on suitable data representations, where either an explicit, or an assumed underlying, ordering of the observations is required, which is combined with assumed exchangeability to enable predictive inference for future observations. As a novel example of NPI, in Section 5 we introduce a variation to $A_{(n)}$ which is suitable for circular data, and we present NPI for such data in Section 6. In Section 7 we discuss NPI in relation to objective Bayesianism. Interestingly, depending on one's views on either topic, one could both advocate and oppose calling NPI a particular form of objective Bayesianism.

2 Hill's assumption $A_{(n)}$

The assumption $A_{(n)}$ was proposed by Hill [44, 45, 46] for prediction in the case of extremely vague a priori knowledge about the form of the underlying distribution. Let real-valued $x_{(1)}, \ldots, x_{(n)}$ be the order statistics of data $x_1, \ldots, x_n$, and let X_i be the corresponding pre-data random quantities, so that the data consist of the realized values, $X_i = x_i$, $i = 1, \ldots, n$. Based on this, $A_{(n)}$ is defined as follows [46]:

1. The observable random quantities $X_1, \ldots, X_n$ are exchangeable. (In the original definition of $A_{(n)}$ [44], exchangeability was not included, allowing slightly more general situations.)

2. Ties have probability 0, so $x_i \neq x_j$, for all $i \neq j$, almost surely. (Generalization to include possible ties is straightforward [45] but leads to more awkward notation.)

3. Given data x_i, $i = 1, \ldots, n$, the probability that the next observation falls in the open interval $I_j = (x_{(j-1)}, x_{(j)})$ is $1/(n+1)$, for each $j = 1, \ldots, n+1$, where we define $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$.

It is clear that $A_{(n)}$ is a post-data assumption related to finite exchangeability [29], see Hill [45] for a detailed presentation and discussion of $A_{(n)}$, and an overview of related work, including important contributions by Dempster [30] and Lane and Sudderth [52]. Hill [46] justified $A_{(n)}$ within the finitely additive Bayesian framework by characterising a corresponding prior process. The strength of the assumption $A_{(n)}$ can best be indicated by citing Hill [45]: 'Let me conclude by observing that $A_{(n)}$ is supported by all of the serious approaches to statistical inference. It is Bayesian, fiducial, and even a confidence/tolerance procedure. It

is simple, coherent, and plausible. It can even be argued, I believe, that $A_{(n)}$ constitutes the fundamental solution to the problem of induction'.

De Finetti's [29] representation theorem provides a Bayesian framework for learning about an underlying parameter, based on infinite exchangeability and using a probability distribution for that parameter. While he relies on the assumption that there is an infinite sequence of random quantities involved, in $A_{(n)}$-based inference we are mostly explicitly interested in a single (or a limited number of) future observation(s). Even more, the Bayesian approach, as justified by De Finetti's [29] important results, explicitly needs a specified prior distribution, and together with the conditional independence of future observations (conditional on an unknown parameter) this adds quite a bit more structure to the data than our $A_{(n)}$-based inferences. Such inferences have a predictive and nonparametric nature, and seem suitable if there is hardly any knowledge about the random quantities of interest, other than the first n observations, or, which may be more realistic, if one explicitly does not want to use such information. This may occur, for example, if one wants to study the (often hidden) effects of additional structural assumptions underlying statistical models or methods. Inferences based on such restricted knowledge have also been called 'low structure inferences' [35] and 'black-box inferences' [52].

The assumption $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest in inference based on such data. However, it does provide bounds for probabilities, by what is essentially an application of De Finetti's 'fundamental theorem of probability' [29], or Walley's concept of 'natural extension' [57]. The theory of interval probability [3, 57, 60, 61, 62] makes it clear that such bounds contain valuable information, both on uncertainty of events and on indeterminacy caused by restricted information.

3 Interval probability

The idea to use interval-valued probabilities dates back at least to the middle of the nineteenth century [10]. Since then, interval probabilities, also known as imprecise probabilities, have been suggested in various areas of statistics. For example, they arise naturally in several approaches to predictive inference such as Dempster's [31] multivalued mappings and Hampel's [43] successful bets, in modelling uncertain knowledge in artificial intelligence [64], in economic decision theory [11], and in robust Neyman-Pearson testing [2, 48]. Furthermore, there is a strong connection to robust Bayesian inference [6, 56]. Recently, there has been increasing activity in this area by researchers from widely varying backgrounds, resulting in a series of conferences [9, 27, 28], (organised by) the Society for Imprecise Probability Theory and Applications[1], and special issues of journals focussing on this topic [8, 25, 26].

Fine and collaborators (e.g. [55, 59]) established a frequentist theory of interval probability.

[1] *www.sipta.org*

Extending De Finetti's [29] theory to interval-valued previsions, Walley [57] provides a rigorous generalization of the concept of probability, based on a behavioural interpretation of subjective lower and upper probabilities as possibly differing maximum buying price and minimum selling price, respectively, for gambles on the event of interest. A formal foundation of interval probability in the spirit of Kolmogorov's axioms, relying on σ-additive classical probabilities, is developed by Weichselberger [62] (see also [60, 61] for some selected aspects). According to Weichselberger [62], an axiomization of interval probability can be achieved by supplementing Kolmogorov's axioms. We briefly present some key aspects of theory of interval probability [62], as relevant to $A_{(n)}$-based inference [3].

Let $(\Omega, \mathcal{A})$ be a measurable space. A set-function $p(\cdot)$ on $\mathcal{A}$ satisfying Kolmogorov's axioms is called a *classical probability*; the set of all classical probabilities on $(\Omega, \mathcal{A})$ is denoted by $\mathcal{K}(\Omega, \mathcal{A})$. A function $P(\cdot)$ on $\mathcal{A}$ is called an *F-probability* with *structure* $\mathcal{M}$, if $P(\cdot)$ is of the form

$$\begin{array}{rcccl} P & : & \mathcal{A} & \to & \left\{[\underline{P}; \overline{P}] \mid 0 \le \underline{P} \le \overline{P} \le 1\right\} \\ & & A & \mapsto & P(A) = \left[\underline{P}(A); \overline{P}(A)\right], \end{array} \tag{1}$$

and

$$\mathcal{M} := \{p(\cdot) \in \mathcal{K}(\Omega, \mathcal{A}) \mid \underline{P}(A) \le p(A) \le \overline{P}(A),\ \forall A \in \mathcal{A}\} \neq \emptyset, \tag{2}$$

and

$$\left.\begin{array}{l} \inf\limits_{p(\cdot)\in\mathcal{M}} p(A) = \underline{P}(A) \\ \sup\limits_{p(\cdot)\in\mathcal{M}} p(A) = \overline{P}(A) \end{array}\right\} \quad \forall A \in \mathcal{A}\,. \tag{3}$$

Throughout this paper, we use notation $P(\cdot)$ and $[\underline{P}(\cdot), \overline{P}(\cdot)]$ for inter-valued assignments, and $p(\cdot)$ for classical probability. Property (2) is a minimal requirement to ensure that $P(\cdot)$ is not contradicting classical probability theory. Property (3) goes beyond this, by requiring a one-to-one correspondence between structure and $P(\cdot)$, guaranteeing that the intervals $[\underline{P}(A), \overline{P}(A)],\ A \in \mathcal{A}$, are not too wide with respect to the structure. Property (3) has been considered by several authors, for instance Huber and Strassen [48], and, if $\mathcal{A}$ is finite, it coincides with Fine's notion of envelopes (e.g. [55, 59]) and with Walley's [57] concept of coherence.

Some consequences of the above definitions are that for every F-probability, $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$ are *conjugated*,

$$\underline{P}(A) = 1 - \overline{P}(A^c), \quad \forall A \in \mathcal{A}\,,$$

which ensures that every F-probability is uniquely characterized by $\underline{P}(\cdot)$, and $\underline{P}(\cdot)$ is super-additive and $\overline{P}(\cdot)$ is subadditive, i.e.,

$$\underline{P}(A \cup B) \ge \underline{P}(A) + \underline{P}(B) \quad \text{and} \quad \overline{P}(A \cup B) \le \overline{P}(A) + \overline{P}(B), \qquad \forall A, B \in \mathcal{A}, A \cap B = \emptyset.$$

In Section 4, an overview of $A_{(n)}$-based interval probability is presented, together with the corresponding nonparametric predictive inferential approach. Augustin and Coolen [3] prove that such interval probability is F-probability. They also provide further insight into these interval probabilities, by showing that they are totally-monotone C-probability [3, 62], but not Choquet capacity [3, 12].

4 Nonparametric predictive inference

It is straightforward to introduce predictive lower and upper probabilities based on the assumption $A_{(n)}$ [3]. Let $\mathcal{B}$ be the Borel σ-field over $\mathbb{R}$. For any element $B \in \mathcal{B}$, set-functions $\underline{P}(\cdot)$ and $\overline{P}(\cdot)$ for the event $X_{n+1} \in B$, based on the intervals $I_1, \ldots, I_{n+1}$ created by n real-valued non-tied observations, and the assumption $A_{(n)}$, are

$$\underline{P}(X_{n+1} \in B) = \frac{1}{n+1}|\{j : I_j \subseteq B\}| \quad (4)$$

$$\overline{P}(X_{n+1} \in B) = \frac{1}{n+1}|\{j : I_j \cap B \neq \emptyset\}|. \quad (5)$$

Throughout this paper, we leave the conditioning on data $x_1, \ldots, x_n$ out of the notation. $\underline{P}(X_{n+1} \in B)$ and $\overline{P}(X_{n+1} \in B)$ can be understood as bounds for the probability for the event $X_{n+1} \in B$, consistent with the probabilities as assigned by $A_{(n)}$. The lower probability $\underline{P}(X_{n+1} \in B)$ is achieved by taking only probability mass into account that is necessarily within B, which is only the case for the probability mass $\frac{1}{n+1}$, per interval I_j, if this interval is completely contained within B. The upper bound $U(\cdot)$ is achieved by taking all the probability mass into account that could possibly be within B, which is the case for the probability mass $\frac{1}{n+1}$, per interval I_j, if the intersection of I_j and B is non-empty. Remark that, in this reasoning, we do allow positive probability masses in points. Augustin and Coolen [3] show that these bounds fit nicely into the framework of interval probability [61, 62]. They are F-probability [3, Thm 2], using all information from $A_{(n)}$ without adding further assumptions. Augustin and Coolen [3] prove that these $A_{(n)}$-based lower and upper probabilities are totally-monotone, which implies (so-to-say 'static') coherence in Walley's [57] sense. According to Walley's generalized betting interpretation for lower and upper probabilities [57], this means that, if we are acting according to our interval probabilities, nobody can place a Dutch book against us at any fixed moment in time, and we accept all sure gains. When using these $A_{(n)}$-based lower and upper probabilities (4) and (5) for statistical inference, they have strong internal consistency properties both from a static and a dynamic point of view [3], where we use 'static' referring to conditioning on a further event for the random quantity X_{n+1}, and 'dynamic' in case of updating on the basis of further observations.

For examples of the use of these lower and upper probabilities (4) and (5) for nonparametric predictive inference (NPI), in statistics and operational research, see [1, 14, 18, 19, 20]. In

these papers, NPI is also compared with more established methods, both via examples using data from the literature, and simulation studies, leading to the conclusion that NPI performs well. Of course, such applications require inferential problems to be explicitly formulated in terms of one future observation. This can be extended to multiple future observations, by sequential conditioning and, after conditioning on the values of the next j observations, assuming the appropriate assumption $A_{(n+j)}$. Such a procedure tends to lead to increased imprecision, that is the difference between corresponding upper and lower probabilities, due to the imprecision involved with conditioning on the next observation(s). We do not address this further here, but we return briefly to multiple future observations in Section 6.4.

Clearly, such inferences require an exchangeability assumption for the observable random quantities. This idea can also be applied in less straightforward situations. For example, in case a data set contains right-censored observations, Coolen and Yan [21, 22, 23] developed a generalization of $A_{(n)}$, called 'right-censoring-$A_{(n)}$', and they illustrated its use in inferential problems, mostly via the corresponding lower and upper survival functions. This provides a predictive alternative to the well-known Product-Limit estimator (PLE) of the survival function for such data, as presented by Kaplan and Meier [51], and it also improves on an earlier attempt to apply $A_{(n)}$ for right-censored data by Berliner and Hill [7]. Coolen and Yan [23] show that the PLE provides a survival function which is always between their lower and upper survival functions, which will be of interest in our later discussion on NPI and objective Bayesianism (Section 7), as the PLE can be interpreted as a generalization of the empirical survival function in case of right-censored data. Coolen-Schrijner and Coolen [24] used right-censoring-$A_{(n)}$ to create highly flexible, fully adaptive methods for age replacement of technical units. Via extensive simulation studies, they found that this method performs well when compared to more established methods for age replacement.

Coolen [16] used $A_{(n)}$ for NPI in case of Bernoulli data, providing lower and upper probabilities for the number of successes in m future trials, based on the number of successes in n observed trials. This was possible by considering the same representation for such Bernoulli data as was used by Bayes [4], namely as balls on a table. However, Coolen replaced the prior distribution (an imaginary 'ball 0' in Bayes' approach), required to link the observations to future outcomes, by $A_{(n)}$. We return to this briefly in Section 7. For any event of interest with regard to m future trials, Coolen's lower and upper probabilities create an interval containing the corresponding observed proportion, again causing such NPI to be consistent with the use of the empirical probabilities.

Recently, Coolen and Augustin [17] have presented a similar approach for multinomial data, using an assumed underlying representation in terms of a probability wheel, where each observation class is assumed to be represented by only a single segment. Their method does not make any assumption on (an upper bound for) the total number of possible categories, if such information is available their method can be adapted accordingly. Again, we will return

to this briefly in Section 7, but we should remark that their lower and upper probabilities are again consistent with the use of the empirical probabilities. The use of an assumed underlying probability wheel representation for multinomial data required introduction of a variation of $A_{(n)}$ suitable for prediction with such a representation. This variation was the same as we discuss in Section 5, where we introduce it directly for NPI for circular data.

Effectively, it has been shown by these methods that NPI can be used whenever one may judge a suitable exchangeability assumption to be appropriate, either directly for the observable random quantities, or via an assumed underlying data representation. Throughout, such inferences are consistent within interval probability theory, in a similar way as discussed above for the $A_{(n)}$-based lower and upper probabilities (4) and (5), and they are particularly suitable in situations where very little information about random quantities of interest is available, or where, perhaps more appropriately, one wishes not to use such information in addition to data.

To illustrate the possibility to develop and apply such NPI to other data situations, we consider circular data in Section 6. First, we present a variation to $A_{(n)}$ which is suitable for nonparametric prediction in case of circular data, which we call 'circular-$A_{(n)}$' and denote by $\circledA_{(n)}$. We illustrate NPI for circular data via 3 examples, using data from the literature. Discussing such data, and generalizations in more dimensions also known as 'directional data', Jammalamadaka and SenGupta [49] call prediction of a future occurrence based on empirical evidence the single most important aim of statistics. They present general methods for predictive inference with directional data, including circular data, but their methods require parametric distributions to be assumed.

Useful introductions to circular data, and the more general concept of directional data, together with overviews of common statistical methodology, are for example the books by Mardia [53], Batschelet [5], and Fisher [33]. Other recent contributions to theory of circular data, and more generally spherical data, include focus on tests of independence of two responses [50] and on asymptotical properties of nonparametric tolerance regions [54].

5 Circular $A_{(n)}$

For circular data, $A_{(n)}$ in its standard form is not suitable, as the data are not represented on the real-line. However, a straightforward variation, linking again to exchangeability of $n+1$ observations, is the assumption that we call *circular-*$A_{(n)}$, and denote by $\circledA_{(n)}$:

Suppose that ordered circular data $x_1 < x_2 < \ldots < x_n$ create n intervals on a circle, namely $C_j = (x_j, x_{j+1})$ for $j = 1, \ldots, n-1$, and $C_n = (x_n, x_1)$. Then we propose assumption $\circledA_{(n)}$ to be that a future random quantity X_{n+1} falls into each of these n intervals with equal

probability, so

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

Notice here that neither the units of the circular data, nor the chosen 0-point on the circle, are relevant. This 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 $\mathcal{A}_{(n)}$ will have the same consistency properties as such inference based on $A_{(n)}$ has [3, 35, 46].

In Section 6, we illustrate NPI for circular data based on $\mathcal{A}_{(n)}$. As we wish not to make further assumptions about the probability mass $1/n$ per interval C_j, such inferences are mostly again in the form of lower and upper probabilities [57, 61, 62].

6 NPI for circular data

In this section, we first present the lower and upper probabilities for a single future observation, based on $\mathcal{A}_{(n)}$ and circular data consisting of n observations. This is a straightforward variation to the $A_{(n)}$-based interval probabilities in Section 4. Thereafter, we illustrate such NPI by considering the comparison of two groups of circular data, and by presenting how this method can deal with grouped circular data. This latter problem is relatively straightforward due to the use of lower and upper probabilities, which means that no additional assumptions are required. We illustrate all these inferences via examples with data taken from the literature. At the end of this section, we briefly comment on the generalization of NPI for circular data to multiple future observations.

6.1 Interval probabilities for a single future observation

The probabilities for X_{n+1} as defined in $\mathcal{A}_{(n)}$ directly lead to lower and upper probabilities for events of the form $X_{n+1} \in B$, with B a segment (or a union of segments) of the circle on which the data are represented, following a similar procedure as presented in Section 4 related to $A_{(n)}$ [3]. Hence, the lower probability $\underline{P}(X_{n+1} \in B)$ is again derived by summing only the probability masses assigned to intervals C_j that are fully contained within B, and the upper probability $\overline{P}(X_{n+1} \in B)$ is derived by summing all probability masses assigned to intervals C_j that have non-empty intersection with B. Here, we explicitly use the fact that we have not added any further assumptions on the distribution of the probability $1/n$ within each interval C_j. It is easy to see that strong consistency results for such inferences, as presented by Augustin and Coolen [3], also hold for these $\mathcal{A}_{(n)}$-based inferences for circular data, in particular considering updating in the light of new observations, and conditioning (as

illustrated briefly in Example 1). These lower and upper probabilities are again F-probability, which follows from the fact that the proofs of Theorem 1 and Theorem 2 in [3] do not use the ordering of the intervals created by the data on the real-line, and therefore these results can be directly adopted for the assumption $\mathcal{A}_{(n)}$ and these corresponding lower and upper probabilities. Notice that, if B is a segment of the circle not containing any observed x_j, then the lower probability of the event $X_{n+1} \in B$ is 0, but the upper probability for this event is equal to $1/n$. We illustrate this procedure using data from an example by Batschelet [5], which were also used by Jammalamadaka and SenGupta [49].

Example 1.
In a city, 21 major traffic accidents were recorded during several days, the times are shown in Table 1 represented as circular data where the circle represents the daily cycle of 24 hours, which is an appropriate representation if one is only interested in the time of day of accidents, and not in the variation of such data over a longer period. We choose to represent these data in 'minutes after midnight', where the full cycle is equal to 1440 minutes. (Jammalamadaka and SenGupta [49] represent these times using angles of the circle, out of 360 degrees, their data values are equal to our data in minutes divided by 4, using angles would make no difference to our inferences.)

56	728	1044
188	808	1088
292	856	1096
436	980	1136
488	1004	1172
600	1024	1252
684	1040	1328

Table 1: Traffic accident data, in minutes (Example 1)

On the basis of these data, assumption $\mathcal{A}_{(21)}$ provides predictive probabilities for X_{22}, the time of the day at which a future accident will occur. It provides precise probabilities $1/21$ for the intervals between two previous observations, and interval probabilities for other events. For example, for the event of a future accident happening 'in the evening', we have lower probability $\underline{P}(X_{22} \in (1080, 1440)) = 5/21 = 0.238$, and upper probability $\overline{P}(X_{22} \in (1080, 1440)) = 7/21 = 0.333$. Alternatively, one could define the traffic 'rush hours' as being between 7 and 9 in the morning, and between 5 and 7 in the evening. The lower and upper probabilities for the event $X_{22} \in (420, 540) \cup (1020, 1140)$ are $6/21 = 0.286$ and $10/21 = 0.476$, respectively, according to $\mathcal{A}_{(21)}$ and these data.

One could also be interested in the conditional probability that such a future accident would take place during the morning rush hours, given that it happens during either the

morning or the evening rush hours. Generally, when the event on which we condition does not have a precise probability assigned to it by $\mathcal{A}_{(n)}$, the conditional lower and upper probabilities follow the intuitive concept [3, 61, 62], and in this case these are equal to

$$\underline{P}(X_{22} \in (420, 540) \,|\, X_{22} \in (420, 540) \cup (1020, 1140)) = \frac{1/21}{1/21 + 7/21} = 1/8 = 0.125,$$

and

$$\overline{P}(X_{22} \in (420, 540) \,|\, X_{22} \in (420, 540) \cup (1020, 1140)) = \frac{3/21}{3/21 + 5/21} = 3/8 = 0.375.$$

6.2 Comparison of two groups of circular data

Statistical inferential problems often involve comparison of two or more independent groups of data. Traditionally, such comparisons are often formulated in terms of hypotheses about equality of characteristics of underlying population distributions, for example equal mean values. From predictive perspective, similar comparisons can be performed, yet these necessarily are restricted to comparisons formulated in terms of future observations from the different groups. For observations on the real-line, such comparisons in NPI have been presented for two groups of data [14] as well as for multiple comparisons [20], for both cases lower and upper probabilities were derived for the event that the next observation from one group exceeds the next observations from the other groups. For circular data, formulation of such comparisons is perhaps less straightforward, due to absence of a unique ordering. However, with a predictive formulation such comparison might naturally be more directly defined in terms of events of particular interest. For example, if one wants to check if there is a reason to believe that the next observation from one group is more likely to belong to a segment B of the circle than the next observation from another group, one can compare the interval probabilities for these events, as based on the appropriate $\mathcal{A}_{(n)}$-assumptions per group, where in particular non-overlapping interval probabilities indicate a relevant difference within theory of interval probability [57, 62]. In Example 2 we illustrate this, using data from an example by Batschelet [5], concerning two groups of data from an experiment on orientation of pigeons. Although we do not address it explicitly, multiple comparisons with circular data can be treated similarly [20].

Example 2.
Two groups of pigeons were carried from their loft near Siena to Rome. During transport, one group ('controls' - for these we add an index 'c' to notation) experienced the natural air along the road, the other group ('experimentals' - index 'e') received only pure air. At the release site, the correct direction back to Siena was 325 degrees, where North was set at 0 degrees with the angle measured clock-wise. The control group consisted of $n_c = 8$ pigeons,

the experimental group of $n_e = 10$ pigeons. The observed vanishing bearings are given in Table 2.

Experimentals		Controls	
4	117	24	247
82	121	153	264
107	131	192	333
109	171	202	
110	186	228	

Table 2: Pigeon data; experimentals and controls, in degrees (Example 2)

For NPI based on these data, we focus on random quantities which represent the vanishing bearings for 'the next pigeon' for each group, i.e. $X_{e,11}$ and $X_{c,9}$, for which we assume $\mathcal{A}_{(10)}$ and $\mathcal{A}_{(8)}$ in combination with the data per group, respectively. For such inferences, we can indeed think in terms of there being actually one more pigeon for both groups, transported and released with the others, but for which the data value has not yet been given, with such pigeons being selected randomly from all the pigeons per group (independent of their actual data value).

For such a study, one may be interested in whether or not the pigeons vanish in 'more or less the right direction', say for this example whether or not the pigeons vanish in directions between 235 and 55 degrees, representing maximum 90 degrees difference from the actual direction. The relevant lower and upper probabilities are:

$$\underline{P}(X_{e,11} \in (235, 55)) = 0 \quad \text{and} \quad \overline{P}(X_{e,11} \in (235, 55)) = \frac{2}{11}$$

for the experimentals, and

$$\underline{P}(X_{c,9} \in (235, 55)) = \frac{3}{11} \quad \text{and} \quad \overline{P}(X_{c,9} \in (235, 55)) = \frac{5}{11}$$

for the controls. These values suggest a clear difference between these two groups of data with regard to the event specified, as $\underline{P}(X_{c,9} \in (235, 55)) > \overline{P}(X_{e,11} \in (235, 55))$. For example, from a subjective point of view [57], this implies that there are prices at which one would be willing to buy the bet for $X_{c,9}$ yet sell the bet for $X_{e,11}$ on this event, implying quite a strong preference for the bet involving the controls for this event. It can be remarked that classical analysis [5] also indicated that these two samples deviate significantly from each other, in the sense that there appears to be a shift in the mean directions. We think that both such analyses can be useful together, as they address quite different questions. In particular when one is really interested in a predictive inferential question, our approach seems more appealing. Also, our approach appears to be more flexible as it does not require problems to be formulated as testable hypotheses.

6.3 Grouped circular data

Statistical data are frequently presented as grouped data, e.g. via histograms, in particular if there are quite a lot of data. Coolen and Yan [21] present the use of NPI for grouped lifetime data, including right-censored observations, as typically appears in the case of life tables used in life insurance. For grouped circular data, NPI is also easily adapted by considering the different possible configurations on the circle which are in agreement with the grouped data information, in which case the lower and upper probabilities tend to correspond to extreme situations with regard to the number of data values that could actually be within a circle segment of interest. We illustrate NPI for grouped circular data using data from an example by Mardia [53].

Example 3.
Table 3 shows the frequencies of vanishing angles (in degrees) of 714 nonmigratory British mallards with 0 as the North (measured clockwise). Mardia [53] presents these data in a variety of ways, including a useful circular histogram which also makes clear that the majority of these birds vanish in directions West to North-West.

Direction	Number	Direction	Number
0-	40	180-	3
20-	22	200-	11
40-	20	220-	22
60-	9	240-	24
80-	6	260-	58
100-	3	280-	136
120-	3	300-	138
140-	1	320-	143
160-	6	340-	69

Table 3: Mallards data; vanishing angles, in degrees (Example 3)

As these data are grouped in intervals of 20 degrees, but we do not have further information about the exact data values, lower and upper probabilities for a random quantity X_{715} representing a future vanishing angle for such a mallard, based on the assumption $\mathcal{A}_{714}$, are derived by taking the extreme configurations of the data per interval with regard to a specified event of interest. For example, suppose we are interested in the event $X_{715} \in (270, 360)$, so such a mallard vanishing in a direction between West and North. Using the probabilities assigned by $\mathcal{A}_{714}$, there are 486 observed values in $(280, 360)$, implying that the probabilities for the 485 intervals between these values are all necessarily also in $(270, 360)$, hence the lower probability is

$$\underline{P}(X_{715} \in (270, 360)) = 485/714 = 0.679.$$

However, the observed values in $(260, 280)$ could also all have fallen in $(270, 280)$, as we have no information on this, and do not wish to add further assumptions, we have to take this into account when calculating the upper probability. So, we now need to consider the extreme case that 544 values may actually be in $(270, 360)$. To derive the upper probability for this interval, we must also think about the probabilities $1/714$ assigned to the intervals which include the values 270 and 360, as without further assumptions these probability masses could actually also be in $(270, 360)$. This leads to upper probability

$$\overline{P}(X_{715} \in (270, 360)) = 545/714 = 0.763.$$

The difference between these upper and lower probabilities is mostly due to the uncertainty about the exact location of the data values on the circle, caused by the grouped manner in which the data are reported. We consider it an attractive feature of this NPI approach, using interval probability, that no further assumptions are required for such inferences. Of course, if one would wish to add further assumptions on the distribution of observations within the intervals used for the data representation, one can immediately see the effect of such further assumptions through the reduction of the imprecision in these interval probabilities.

6.4 Multiple future observations

While $\mathcal{A}_{(n)}$ is an assumption that provides a predictive probability for only a single future observation, it can be extended to m future observations, represented by random quantities $X_{n+1}, \ldots, X_{n+m}$, by effectively assuming simultaneously $\mathcal{A}_{(n)}, \mathcal{A}_{(n+1)}, \ldots, \mathcal{A}_{(n+m-1)}$, similarly as was done by Hill [44] for $A_{(n)}$, and as was also derived by Dempster [30] from slightly different perspective. Let S_j be the number out of these m future observations that fall into C_j, then these combined assumptions lead to probabilities

$$p(\bigcap_{j=1}^{n} \{S_j = s_j\}) = \binom{n+m-1}{m}^{-1},$$

where s_j, for $j = 1, \ldots, n$, are any non-negative integers with $\sum_{j=1}^{n} s_j = m$. These probabilities allow NPI for multiple future observations simultaneously, and the results in the previous sections can fairly straightforwardly be generalized to m future observations, where however the differences between upper and lower probabilities tend to increase with m, see Coolen [16] for a similar study of NPI for m future observations in case of Bernoulli random quantities. To calculate lower and upper probabilities for an event of interest, e.g. for the event that, based on n observations, at most b out of $m \geq 2$ future observations belong to a segment B of the circle, one needs to count the number of different sets $\{s_1, \ldots, s_n\}$, with $\sum_{j=1}^{n} s_j = m$, for which not more than b future values can be in B, leading to the lower probability for this event, and the number for which it is possible that at most b future values can belong to B, leading to the upper probability. More complex events of interest may require sophisticated counting methods, but the basic principle remains the same.

7 NPI and objective Bayesianism

Williamson [63] presents a detailed overview of objective Bayesianism, and its challenges. While Bayesian inference, based on subjective interpretation of probabilities and decision making [29, 34], requires uncertainty quantification via probabilities satisfying axioms of probability, objective Bayesianism imposes two further norms [63]:

(EN) Empirical: An agent's knowledge of the world should constrain her degrees of belief. Thus if one knows that a coin is symmetrical and has yielded heads roughly half the time, then one's degree of belief that it will yield heads on the next throw should be roughly 1/2.

(LN) Logical: An agent's degrees of belief should also be fixed by her lack of knowledge of the world. If the agent knows nothing about an experiment except that it has two possible outcomes, then she should award degree of belief 1/2 to each outcome.

Walley [57] discusses 'objectivity' in detail, and convincingly advocates that classical (precise) probability is too restrictive for achieving objectivity. His arguments even go against the possibility of objective inference. Nevertheless, he aims at achieving 'objective inferential' methods, particularly his imprecise Dirichlet model (IDM) [58] is suggested to be non-subjective. Whereas we mostly support Walley's arguments, we acknowledge objective inference as an ideal in science. However, we strongly feel that this ideal is not fully achievable, and do not wish to make any claims on the objectivity of NPI-based lower and upper probabilities.

It is interesting to consider the above norms for objective Bayesianism in more detail. The norm (EN) is explicitly phrased in a predictive manner, fitting with NPI. When restricted to precise probability, it may appear to be hard to disagree with (EN). However, the word 'roughly' appears both in relation to the observations, and in relation to the apparently logically imposed predictive probability, in rather a vague manner. For example, suppose that 498 out of 1,000 tosses of such a coin give 'heads' (H) as outcome, would one require the predictive probability of the next toss to give H to be equal to 0.498, or perhaps equal to 0.500? If one feels that such data would support strong further knowledge of this particular coin, and the way it is tossed, the latter may be more logical. What, however, if the coin is spun on its edge instead of tossed? Student experiments with Dutch (pre-Euro) coins showed that coins, appearing symmetrical to the naked eye, did not have a tendency to fall heads up in about half the spins. A possible explanation of this behaviour was in the production process of coins, in particular the way they are cut out of large sheets of metal. Suppose that such a coin leads to 416 H out of 1,000 spins, should the predictive probability for H on the next spin of the same coin be 0.416? And, what should the probability distributions be for the number of H's in the next 100, 1,000, or even 100,000 spins?

Perhaps we may loosely interpret (EN) in the following manner:

(EN') Objective inferences should not disagree with empirical evidence.

Of course, (EN') is equally vague, in particular with regard to the word 'disagree'. When restricted to precise probability, one could interpret (EN') as stating that a predictive probability should equal an observed frequency of a particular outcome category, or, for real-valued observations, that the predictive distribution function should equal the empirical distribution function. The latter, of course, may cause difficulties due to the discrete nature of empirical distribution functions. The first seems acceptable, but perhaps more so if we have many observations than if one only has few observations. Also, one may object to predictive probabilities being 0 for not yet observed outcomes, both for small and larger data sets. We discuss this in more detail in Section 7.1.

The intervals created by our lower and upper probabilities, as based on $A_{(n)}$ and its variations as discussed earlier in this paper, always include the empirical probabilities according to the data set used. Hence, such NPI never disagrees with empirical evidence, when we explicitly restrict 'empirical evidence' to the n observations. Although this property is not as strong as one may like, as e.g. vacuous interval probabilities (i.e. $\underline{P}(\cdot) = 0$ and $\overline{P}(\cdot) = 1$) also trivially satisfy this property, it should be emphasized that NPI is far from vacuous, with imprecision decreasing as a function of n.

The norm (LN), which is also formulated in a predictive manner, also seems attractive from a classical probability perspective, as it would be difficult to advocate a different norm. However, an important question is whether or not one would impose the agent to act accordingly. Perhaps even more worryingly, and in line with increasingly popular use of uncertainty quantifications, is the question whether or not one should include such a probability of 1/2 in an 'expert system', without further quantification (or description) of the strength of evidence on which this probability is based. This norm (LN) reflects a strong symmetry assumption, probably caused by the restrictive nature of precise probability [57]. For use of uncertainty quantifications in expert systems, it may be the case that a single number representation is too restrictive. Bayesian statisticians may argue that one can reflect such strength of evidence via higher-order probabilities. From subjective perspective, they may misunderstand De Finetti's [29] case for fair prices of gambles as the concept on which his justification of Bayesian inference relies: he clearly states that for any gamble one must have a fair price (whether or not one wishes to report this, either fairly or not at all, is a different matter). Hence, at worst higher-order probabilities do not fit in this concept, a defendable point of view for any probabilities on non-observable events (how to settle the gamble?), and at best the use of higher-order probabilities significantly complicates storing and using information in expert systems. And, let us not forget that, if interest is explicitly in the next observation,

any higher-order probabilities are integrated out to derive the required marginal probability, again leading to a single value which fails to represent the strength of empirical evidence. Walley [57] discusses such symmetry assumptions in detail.

Perhaps we may loosely interpret (LN) in the following manner:

(LN') If one has no information suggesting that one possible outcome is more likely than another, then this should be reflected by identical uncertainty quantifications for these outcomes.

Our formulation (LN') avoids mentioning explicitly the number of possible outcomes, which is probably required to be known for (LN) to be applicable. We discuss this further in Section 7.1. By the use of lower and upper probabilities, our nonparametric predictive inferences are relatively straightforward to use in expert systems, far easier than, for example, if higher-order probability distributions would need to be stored. In the most extreme situation, where one has no data whatsoever, and where, as explicitly assumed for NPI, one does not wish to use further information in a subjective manner, we can have vacuous interval probabilities for all but trivial events involving the next observation(s). Indeed, we could not even apply $A_{(n)}$, nor its variations, in a sensible manner for $n = 0$, and our NPI framework does not define any lower and upper probabilities in case $n = 0$. Nevertheless, the only values that would be consistent with NPI, if $n = 0$, are the vacuous interval probabilities.

Under classical probability, (EN) and (LN) may require to be prioritised, for example because conflict may appear, see Williamson [63] for a detailed discussion. In addition, there is wide scope with regard to application of (LN), where for example a variety of information measures can be used. The use of lower and upper probabilities simplifies such matters enormously, although one still requires assumptions to support a particular choice of values of such interval probabilities. We suggest that NPI presents such particular values, based on clearly stated assumptions with regard to a data representation (either directly on observables, or assumed for an underlying data representation) and post-data exchangeability. The NPI-based lower and upper probabilities can be said to be 'sensible', in the sense that the empirical probabilities are always in the intervals created by the corresponding NPI interval probabilities, and that the length of such intervals decreases as a function of n (leading to precise probabilities for $n \to \infty$). Furthermore, the NPI-based interval probabilities can be said to be 'reasonable' from practical perspective, as they are not to conservative (so not creating intervals that are too wide) when based on medium sized data sets, as typically appear in practical applications, see the examples in Section 6. A further intuitively attractive property, as briefly mentioned for Bernoulli data in Section 7.1, is that the difference between corresponding NPI-based upper and lower probabilities tends to increase as function of the number of future observations specified in the event of interest. Hence, NPI on the basis of 1,000 observations from spinning a coin leads to very little imprecision when considering a

single future observation, but quite much imprecision when considering 100,000 future observations [16]. Walley explored the natural relation between imprecision on an event A, defined as $\Delta(A) = \overline{P}(A) - \underline{P}(A)$, and information, and suggested the intuitively attractive information measure $i(A) = \Delta^{-1}(A) - 1$, so $\Delta(A)$ is decreasing as function of $i(A)$, classical probability ($\Delta(A) = 0$) corresponds to $i(A) = \infty$, and vacuous interval probability ($\Delta(A) = 1$) corresponds to $i(A) = 0$. When using this information measure for NPI, $i(A)$ is often quite naturally related to n. Again, we do not wish to make strong claims about this argument as a proof of suitability of either NPI or $i(A)$, but it certainly suggests that the relation between imprecision and information measures is an interesting topic for future research. Coolen [13] explored the use of such a relation explicitly in order to control imprecision upon updating, in a parametric model context close to robust Bayesian theory [6, 56].

With any inferential method, it is important to clarify its possible use in applications. Our position here can be formulated as follows. We advocate the use of NPI as objective input in subjective decision processes. Care should be taken that we do not state that the NPI-based interval probabilities should be used as such, but the 'input' suggested is of the nature 'NPI, with this particular model, data representation, and post-data exchangeability assumption, implies these lower and upper predictive probabilities'. Although we appreciate the ideal of achieving an objective inferential theory, we do not foresee a possible theory which convincingly excludes all subjective judgements. In particular, predictive inference seems to require subjective input with regard to judgement of relevance of data, and some form of 'exchangeability-type judgement' to link data to predictions for future observations. In addition, any model assumptions seem to require at least some subjective elements, for example with regard to the appropriate level of detail.

With regard to NPI, we wish to emphasize the explicitly post-data nature of the assumption $A_{(n)}$ and its variations. In the standard Bayesian approach to statistics, it seems often to be taken for granted that one can fully assess all relevant aspects of an inferential problem, all possible related data structures, and all possible observations, at the prior stage. Goldstein [36, 37, 38, 39, 40] addresses such issues in great detail, making clear that the main philosophical issue with regard to Bayesian statistics is the interpretation of the posterior distribution: there is no logical requirement to use the conditional probability derived at the prior stage as one's actual subjective probability after further data has become available. Indeed, when using a parametric model one may often decide on summary statistics at the prior stage, whereas study of data might then suggest further information from the data to be important. In our NPI setting, one should clearly judge the post-data exchangeability assumption in the light of the data. For example, in case of Bernoulli data, suppose that the outcomes of 1,000 spins of a coin are such that the first 416 spins gave heads, followed by 584 tails. One may then not wish to use the post-data exchangeability assumption for spin 1,001, which effectively assumes that all 1,001 spins involved where similar processes. It is

interesting to remark that Goldstein's work, as mentioned above, on the foundations of prior and posterior inferences, supports the elegant and powerful theory of Bayes linear analysis [41, 42], which is an explicitly subjective approach to statistics and decision making, in line with De Finetti's theory [29]. Bayes linear analysis uses prevision ('expectation') as the primary concept for quantification of uncertainty. Generalizing Bayes linear methods to allow interval-valued prevision is an interesting topic for future research.

An often stated 'disadvantage' of the use of lower and upper probabilities, for statistical inference and decision making, is the possibility of 'indecision', see Walley [57] for a detailed discussion. By emphasizing the difference between the actual decisions or inferences, and the NPI assumptions and results as input to this process, we hope our position is clear: we strongly prefer to indicate all 'reasonable decisions or inferences' in line with the evidence, over the suggested clarity of a single 'optimal' decision or inference without explicit opportunity to appreciate the way in which this is influenced by data and by assumptions, the latter often hidden and used for mathematical simplicity. One of the author's reasons to study and develop NPI has always been a strong wish to understand precisely those influences of data and assumptions on the results of statistical analyses. We suggest to use a variety of statistical models and inferential approaches in parallel, whenever possible. If inferences based on other models and approaches strongly conflict with NPI results, we would wish to study the assumptions underlying the other inferences in detail. Of course, in many situations, particularly when sufficient data are available, one would normally expect the actual inferences, based on different statistical approaches, to be roughly in agreement. And, if in such cases the event of interest enables fairly robust inference [6, 47, 56], the outcomes of most established approaches and NPI will be roughly in agreement, which can provide strong confidence in the resulting decision or action.

7.1 NPI for Bernoulli and multinomial data

We briefly present the main results on NPI for Bernoulli [16] and multinomial data [17], which highlight some of the comments in the above discussion.

For Bernoulli data, Coolen [16] presents NPI-based lower and upper probabilities for general events of interest concerning m future observations ('trials') based on n observations, using an underlying assumed data representation similar to the thought experiment used by Bayes [4], together with the appropriate $A_{(n)}$ assumptions. Here, we only consider the simple event of r successes occurring in m future trials, given s successes have occurred in n trials, which we denote by $(m, r)|(n, s)$, where only the number of future successes r is a random quantity as we assume m to be fixed. The NPI-based lower and upper probabilities presented

by Coolen [16] for this event, are (for $0 \leq s \leq n$ and $0 \leq r \leq m$)

$$\underline{P}((m,r)|(n,s)) = \begin{cases} \frac{\binom{s+r-1}{r}\binom{n-s+m-r-1}{m-r}}{\binom{n+m}{n}} & \text{for } 0 < s < n, \\ \frac{n}{n+m} & \text{for } (s = 0 \text{ and } r = 0) \text{ or } (s = n \text{ and } r = m), \\ 0 & \text{for } (s = 0 \text{ and } r > 0) \text{ or } (s = n \text{ and } r < m), \end{cases}$$

and

$$\overline{P}((m,r)|(n,s)) \; = \; \frac{\binom{m}{r}\binom{n}{s}}{\binom{n+m}{s+r}}.$$

For the special case of a single future observation ($m = 1$), the lower and upper probabilities of a success ($r = 1$) are $\frac{s}{n+1}$ and $\frac{s+1}{n+1}$, respectively. We refer to Coolen [16] for further corresponding results, in particular lower and upper probabilities for r in subsets of $\{0, \ldots, m\}$. For any event of interest, and for any data set and choice of m, the interval created by lower and upper probabilities contains the corresponding empirical probability based on the observed proportion of successes. Also, it is easy to confirm that imprecision decreases as function of n, and increases as function of m, as long as the proportions of successes (both in the data and in the event of interest for the future trials) remain about constant.

Recently, Coolen and Augustin [17] have developed an NPI approach for multinomial data, explicitly assuming no further knowledge about the number of possible observation categories in addition to the data representation, and also no natural ordering or other relations between such categories. They present lower and upper predictive probabilities for a single future observation, Y_{n+1}, based on data consisting 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 $1 \leq k \leq n$. However, adding further specifically defined categories to this data description, to which no observations belong, does not influence any of their inferences. To derive results for all possible events of interest for the next observation, notation is introduced for new, as yet unseen, categories. It is important to distinguish between *D*efined *N*ew categories, of which one needs 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 any as yet *U*nobserved *N*ew observation, which we denote by UN and includes observations in the categories DN_i. By allowing $l \geq 0$ and $0 \leq r \leq k$ in the notation, one can define two types of events that comprise the most generally formulated events that need to be considered for Y_{n+1} in this multinomial setting. These two general events are

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

and

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

Coolen and Augustin [17] present the following NPI-based lower and upper probabilities for these events. For the first of these general events, the lower probability is

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

and the corresponding upper probability is

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

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

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

and the corresponding upper probability is

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

It is easy to confirm that these lower and upper probabilities satisfy the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$ for all events A considered for the next observation, for a chosen data representation. In fact, they are again F-probability [17], and hence strongly consistent. The empirical probability corresponding to an event A, that is the relative frequency of A in the data, is always bounded by these $\underline{P}(A)$ and $\overline{P}(A)$. For detailed illustration of these lower and upper probabilities we refer to [17], but it is interesting to consider these for the event that the next observation belongs to an as yet unseen category, which is one of the possibly more controversial inferences in any statistical theory for which objectivity is claimed. If all n observations belong to one category, then

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

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

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

but

$$[\underline{P}; \overline{P}](Y_{n+1} \in DN_i) = [0; 1/n].$$

Also, the lower probability for the event that the next observation belongs to the same category as a previous observation, only becomes positive if two previous observations belonged to that category. Of course, one may consider it more likely that the next observation belongs to a category already once observed than to a new category, this is reflected via different upper probabilities for these events. For further illustration and discussion of these lower and upper probabilities we refer to [17], where in particular a detailed comparison is given of these NPI-based inferences with Walley's Imprecise Dirichlet Model (IDM) [58], which was proposed for similar inferences. In Walley's model, the lower probability for the next observation to belong to a once observed category is positive, and can actually be pretty large (depending on the choice of a further parameter value that must be specified subjectively, and independently of the data), which was mentioned as a possible disadvantage by several discussants of Walley's paper [15, 58]. In addition, Walley's IDM does not distinguish between UN and DN_i, and the IDM interval probabilities corresponding to the above events, with the extreme data situations of either all observations belonging to the same category or all belonging to different categories, are equal, which was also considered to be a disadvantage of the IDM, both by discussants and by Walley himself [58].

Clearly, these NPI lower and upper probabilities for Bernoulli and multinomial data are in line with the reformulated norms (EN') and (LN') above. In particular, for these predictive inferences based on multinomial data, the actual description of categories is irrelevant once a particular data representation is chosen. However, such inferences do depend on the choice of data representation, where imprecision tends to increase with increased detail of data representation [17].

7.2 NPI: Bayesian or not? Objective or not?

To judge any possible role for NPI in a theory of objective Bayesianism, important questions are whether or not NPI is Bayesian, and whether or not it is objective. We will discuss these issues, without however resolving them.

One could argue both in favour of, and against, a claim that NPI is Bayesian. Hill [44, 45, 46] clearly considered $A_{(n)}$ as a suitable basis for nonparametric Bayesian statistics, and proved its consistency by developing a prior process (under finite additivity) that results in the $A_{(n)}$ assignments as posterior predictive probabilities for the next observation. However, as clearly stated by Hill [45], $A_{(n)}$ is also a 'frequentist procedure'. We adhere more strongly to this latter point of view. Generally, though, both for frequentist and Bayesian approaches, statisticians are often happy to assume exchangeability at the prior stage. Once data are used in combination with model assumptions, exchangeability no longer holds 'post-data' due to the influence of modelling assumptions, which effectively are based on mostly subjective input added to the information from the data. Hence, we consider $A_{(n)}$, and its variations such as $\mathcal{A}_{(n)}$ as presented in this paper, as a natural basis for inference in case one wishes to reduce

subjective input.

When using interval-valued uncertainty quantifications, it is clear that conditioning and updating are explicitly different actions [3, 32]. Hence, it may appear difficult to call updating in NPI, which simply means extending the data set by including a further j observations, and adopting, if one wishes, $A_{(n+j)}$ for further NPI, 'Bayesian'. However, as our NPI-based inferences are strongly internally consistent [3], statically coherent in Walley's sense [3, 57], and also consistent upon updating [3], we think that, in principle, one could argue that they may be considered as 'Bayesian'. For example, if one would bet according to the interpretation of these lower and upper probabilities as prices for which gambles are desirable [57], one cannot be made a sure loser at any particular moment in time.

Our NPI does not require a prior probability, or a set of prior probabilities [6, 57], so from this perspective one may not wish to consider NPI to be Bayesian. Of course, with regard to objective Bayesianism, this nicely avoids the difficult issue of selecting a 'non-informative' prior, or even defining what this means. For example, even for the circular data inferences discussed in Section 6, a variety of priors could be advocated as being 'non-informative', and this is more complex when considering random quantities on the real-line. Also our explicit emphasis on $A_{(n)}$ as a post-data assumption is not in line with Bayesian foundations, where such judgements are required, and only allowed, at the prior stage.

The second question, whether or not NPI is objective, we leave for the reader to judge. This is in line with our earlier claim that we consider NPI to be objective input for decision processes which necessarily depend on subjectivity, where the objectivity must be interpreted with regard to the entire context including model and data representations, including the $A_{(n)}$-type post-data exchangeability assumption used, so not only with regard to the resulting interval probabilities. It would be interesting to compare our NPI for circular data, using the examples in Section 6, with 'objective Bayesian' approaches to the same inferential problems. We would be extremely surprised if any Bayesian approach, that is claimed to be 'objective', would result in predictive probabilities and corresponding inferences which differ noticeable from the NPI results presented in this paper.

Acknowledgements

Sections 2 to 4 of this paper are largely based on similar presentation in my paper [3] with Thomas Augustin (Ludwig-Maximilians University, Munich). Our long-term collaboration on NPI and interval probability has been invaluable for development of this approach.

References

[1] Arts, G.R.J., Coolen, F.P.A. and van der Laan, P. (2004). Nonparametric predictive inference in statistical process control. *Quality Technology and Quantitative Management*, **1**, 201-216.

[2] Augustin, T. (2002). Neyman-Pearson testing under interval probability by globally least favorable pairs; Reviewing Huber-Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, **105**, 149-173.

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

[4] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370-418 *and* **54**, 296-325. Reproduced in: Press, S.J. (1989). *Bayesian Statistics*. Wiley, New York, 185-217.

[5] Batschelet, E. (1981). *Circular Statistics in Biology*. Academic Press, London.

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

[7] Berliner, L.M. and Hill, B.M. (1988). Bayesian nonparametric survival analysis (with discussion). *Journal of the American Statistical Association*, **83**, 772-784.

[8] Bernard, J.M. (Ed.) (2002). Special issue on imprecise probabilities and their applications. *Journal of Statistical Planning and Inference*, **105**, issue 1.

[9] Bernard, J.M., Seidenfeld, T. and Zaffalon, M. (Eds.) (2003). *ISIPTA'03 - Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*. Lugano University. Proceedings in Informatics 18, Carlton Scientific.

[10] Boole, G. (1854). *An Investigation of the Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities*. Walton & Maberley, London. (Reprinted: Dover, New York, 1951).

[11] Chateauneuf, A., Cohen, M. and Meilijson, I. (1997). New tools to better model behavior under risk and uncertainty: an overview. *Finance*, **18**, 25-46.

[12] Choquet, G. (1954). Theory of capacities. *Annals of the Institute Fourier Grenoble*, **5**, 131-295.

[13] Coolen, F.P.A. (1994). On Bernoulli experiments with imprecise prior probabilities. *The Statistician*, **43**, 155-167.

[14] Coolen, F.P.A. (1996a). Comparing two populations based on low stochastic structure assumptions. *Statistics & Probability Letters*, **29**, 297-305.

[15] Coolen, F.P.A. (1996b). Contribution to discussion of [58]. *Journal of the Royal Statistical Society B*, **58**, 43.

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

[17] Coolen, F.P.A. and Augustin, T. (2005). Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. To appear in *Proceedings ISIPTA'05, Carnegie Mellon, July 2005.*

[18] Coolen, F.P.A. and Coolen-Schrijner, P. (2000). Condition monitoring: a new perspective. *Journal of the Operational Research Society*, **51**, 311-319.

[19] Coolen, F.P.A. and Coolen-Schrijner, P. (2003). A nonparametric predictive method for queues. *European Journal of Operational Research*, **145**, 425-442.

[20] Coolen, F.P.A. and van der Laan, P. (2001). Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference*, **98**, 185-203.

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

[22] Coolen, F.P.A. and Yan, K.J. (2003b). Nonparametric predictive comparison of two groups of lifetime data. In [9], 148-161.

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

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

[25] Cozman, F. and Moral, S. (Eds.) (2000). Special volume on imprecise probabilities. *International Journal of Approximate Reasoning*, **24**, 121-299.

[26] de Cooman, G. (Ed.) (2000). Special issue on imprecise probabilities. *Risk, Decision and Policy*, **5**, 107-181.

[27] de Cooman, G., Cozman, F.G., Moral, S. and Walley, P. (Eds.) (1999). *ISIPTA'99: Proceedings of the First International Symposium on Imprecise Probabilities and their Applications.* University of Ghent.

[28] de Cooman, G., Fine, T.L., Moral, S. and Seidenfeld, T. (Eds.) (2001). *ISIPTA'01: Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications.* Cornell University. Shaker Maastricht.

[29] De Finetti, B. (1974). *Theory of Probability* (2 volumes). Wiley, London.

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

[31] Dempster, A.P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, **38**, 325-339.

[32] Dubois, D. and Prade, H. (1994). Focusing versus updating in belief function theory. *Advances in the Dempster-Shafer Theory of Evidence*, Yager, *et al* (Eds.). Wiley, New York, 71-95.

[33] Fisher, N.I. (1993). *Statistical Analysis of Circular Data.* Cambridge University Press.

[34] French, S. and Rios Insua, D. (2000). *Statistical Decision Theory.* Arnold, London.

[35] Geisser, S. (1993). *Predictive Inference: an Introduction.* Chapman and Hall, London.

[36] Goldstein, M. (1983). The prevision of a prevision. *Journal of the American Statistical Association*, **78**, 817-819.

[37] Goldstein, M. (1985). Temporal coherence. *Bayesian Statistics 2*, J.M. Bernardo, *et al.* (eds.). North Holland, 231-248.

[38] Goldstein, M. (1986). Exchangeable belief structures. *Journal of the American Statistical Association*, **81**, 971-976.

[39] Goldstein, M. (1994). Revising exchangeable beliefs: subjectivist foundations for the inductive argument. *Aspects of Uncertainty*, Freeman and Smith (Eds.). Wiley, Chichester, 201-222.

[40] Goldstein, M. (1996). Prior inferences for posterior judgements. *Proceedings of the 10th International Congress of Logic, Methodology and Phylosophy of Science*, Chiara, *et al* (Eds.). Kluwer.

[41] Goldstein, M. (1999). Bayes linear analysis. *Encyclopaedia of Statistical Sciences*, update volume 3, Kotz, *et al.* (Eds.). Wiley, New York, 29-34.

[42] Goldstein, M. and Wooff, D.A. (1995). Bayes linear computation: concepts, implementation and programs. *Statistics and Computing*, **5**, 327-341.

[43] Hampel, F. (1997). What can the foundations discussion contribute to data analysis? And what may be some of the future directions in robust methods and data analysis? *Journal of Statistical Planning and Inference*, **57**, 7-19.

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

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

[46] Hill, B.M. (1993). Parametric models for A_n: splitting processes and mixtures. *Journal of the Royal Statistical Society B* **55**, 423-433.

[47] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

[48] Huber, P.J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *Annals of Statistics*, **1**, 251-263. (Correction: 2, 223-224.)

[49] Jammalamadaka, S.R. and SenGupta, A. (1998). Predictive inference for directional data. *Statistics & Probability Letters*, **40**, 247-257.

[50] Johnson, R.A. and Shieh, G.S. (2002). On tests of independence for spherical data-invariance and centering. *Statistics & Probability Letters*, **57**, 327-335.

[51] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

[52] Lane, D.A. and Sudderth, W.D. (1984). Coherent predictive inference. *Sankhyā Series A*, **46**, 166-185.

[53] Mardia, K.V. (1972). *Statistics of Directional Data*. Academic Press, London.

[54] Mushkudiani, N.A. (2002). Small nonparametric tolerance regions for directional data. *Journal of Statistical Planning and Inference*, **100**, 67-80.

[55] Papamarcou, A. and Fine, T.L. (1991). Unstable collectives and envelopes of probability measures. *Annals of Probability*, **19**, 893-906.

[56] Rios Insua, D. and Ruggeri, F. (Eds.) (2000). *Robust Bayesian Analysis*. Lecture Notes in Statistics 152, Springer, New York.

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

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

[59] Walley, P. and Fine, T.L. (1982). Towards a frequentist theory of upper and lower probability. *Annals of Statistics*, **10**, 741-761.

[60] Weichselberger, K. (1995). Axiomatic foundations of the theory of interval-probability. *Proceedings of the Second Gauß Symposion, Section B*. Mammitzsch and Schneeweiß (Eds.). De Gruyter, Berlin, 47-64.

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

[62] Weichselberger, K. (2001). *Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als Umfassendes Konzept*. Physika, Heidelberg (in German).

[63] Williamson, J. (2004). Philosophies of probability: objective Bayesianism and its challenges. *Handbook of the Philosophy of Mathematics*, Volume 9 of the *Handbook of the Philosophy of Science*, A. Irvine (ed.). Elsevier.

[64] Yager, R.R., Fedrizzi, M. and Kacprzyk, J. (Eds.) (1994). *Advances in the Dempster-Shafer Theory of Evidence*. Wiley, New York.