Multinomial Nonparametric Predictive Inference:

Selection, Classification and Subcategory Data

Rebecca M. Baker

A Thesis presented for the degree of
Doctor of Philosophy

Statistics and Probability Research Group
Department of Mathematical Sciences
University of Durham
England

March 2010

Multinomial Nonparametric Predictive Inference: Selection, Classification and Subcategory Data

Rebecca M. Baker

Submitted for the degree of Doctor of Philosophy
March 2010

Abstract

In probability and statistics, uncertainty is usually quantified using single-valued probabilities satisfying Kolmogorov's axioms. Generalisation of classical probability theory leads to various less restrictive representations of uncertainty which are collectively referred to as imprecise probability. Several approaches to statistical inference using imprecise probability have been suggested, one of which is nonparametric predictive inference (NPI). The multinomial NPI model was recently proposed [14, 17], which quantifies uncertainty in terms of lower and upper probabilities. It has several advantages, one being the facility to handle multinomial data sets with unknown numbers of possible outcomes. The model gives inferences about a single future observation.

This thesis comprises new theoretical developments and applications of the multinomial NPI model. The model is applied to selection problems, for which multiple future observations are also considered. This is the first time inferences about multiple future observations have been presented for the multinomial NPI model. Applications of NPI to classification are also considered and a method is presented for building classification trees using the maximum entropy distribution consistent with the multinomial NPI model. Two algorithms, one approximate and one exact, are proposed for finding this distribution. Finally, a new NPI model is developed for the case of multinomial data with subcategories and several properties of this model are proven.

Declaration

The work in this thesis is based on research carried out at the Statistics and Probability Research Group, Department of Mathematical Sciences, University of Durham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

First, I would like to say thank you to Frank. I am so grateful for all your help and advice: you have been the best supervisor ever but you have also listened to me and looked after me like a second father. I would also like to thank Pauline. She inspired and encouraged me to undertake my PhD and I wish I could tell her how much I appreciate that. Thanks also to Thomas and Carolin in Munich and to Joaquin and Andres in Granada for interesting discussions and collaborations that contributed greatly to this thesis.

I would like to thank my Mum and Dad for their constant support and love and for always being proud of me. Thank you to Alex, for everything. To Danny, Ric, Jonathan, Nathan and Ian: thanks for keeping an eye on me throughout the various stages of my life in Durham (we've had lots of fun and I'll miss you all) and thanks for helpful discussions about imprecise probability and computer-related issues. Thanks to Emma, Nina and Sarah for equally important discussions about less mathematical topics, and for generally being amazing friends.

To all of these people, thank you for your part in the completion of this thesis and thank you for helping to make my time in Durham unforgettable.

Contents

Chapter 1

Introduction

In probability and statistics, uncertainty is usually quantified using classical probabilities satisfying Kolmogorov's axioms. Generalisation of classical probability theory leads to various less restrictive representations of uncertainty which are collectively referred to as imprecise probability. The concept of imprecise probability has a long history, but in the last twenty years there has been much new research in this field. In 2002, the Society for Imprecise Probability: Theories and Applications (SIPTA) was founded. This society organises conferences, workshops and summer schools in addition to providing access to information and publications about imprecise probability through its webpage (www.sipta.org) and has succeeded in raising awareness of the potential of imprecise probability. Many new methods for uncertainty quantification have been proposed and clear advantages over the classical theory of probability have been shown, indicating that the ongoing development of imprecise probability theories and applications is an exciting and important area of research.

The ever-increasing interest in imprecise probability has led to various new approaches to statistical inference, one of which is nonparametric predictive inference (NPI) [6, 14, 17]. NPI is an inferential framework that uses lower and upper probabilities and has attractive properties from the perspective of interval probability theory [43]. Many applications of NPI have already been presented in the literature and its development is gathering strong momentum with regard to

new methodology for practical applications.

In this thesis, we present several new developments to the theory of NPI for multinomial data. Inferences about multiple future observations are considered for the first time and a new NPI model is presented for multinomial data with subcategories. We also consider methods based on this theory for selection problems and for classification problems, providing an important and innovative step in the development of practical applications of NPI for multinomial data.

In Chapter 2, the background literature relevant to this thesis is summarised. We give a general introduction to imprecise probability and an overview of NPI. This is followed by a summary of classification methods from the literature which are relevant to Chapter 4.

Chapter 3 begins with a summary of relevant selection methods from the literature. We present NPI-based selection methods using a single future observation, then we derive NPI lower and upper probabilities for events involving multiple future observations from a multinomial data set and we show how these can be applied to the problems of category selection and subset selection for multinomial data. The extension of NPI to inferences about multiple future observations is an ongoing topic to which an important contribution is made here. The results of Chapter 3 were presented at the 2009 International Symposium on Imprecise Probability: Theories and Applications [7] and a journal paper on this work is currently under review [8].

In Chapter 4, we present the use of NPI for building classification trees. We test our methods on forty data sets and we use the results to compare NPI-based classifiers with other methods from the literature. Much of the work in Chapter 4 was carried out in collaboration with Joaquin Abellan and Andres Masegosa of the University of Granada, and a paper focusing on the theoretical results of this chapter is currently under review [3]. The development of NPI-based algorithms

for classification trees is a significant contribution with many potential applications and a further paper on this subject is in preparation.

In Chapter 5, we present a new model which is an extension of the multinomial NPI model to the case of multinomial data with subcategories. We derive NPI lower and upper probabilities for all events of interest involving a single future observation and we present several properties of the model. We also consider the application of the model to classification trees. The development of this model is a considerable addition to the theory of NPI and the flexibility of the inferences in the sense that observations can be represented as subcategories or main categories makes the model widely applicable to practical problems. A paper on this model is in preparation.

There are many interesting opportunities to extend the research presented in this thesis, as discussed in the final sections of each of Chapters 3 to 5 and in Chapter 6.

Chapter 2

Preliminaries

In this chapter we summarise the main theories and concepts from the literature that provide relevant background information for the topics considered in this thesis. An introduction to imprecise probability, with emphasis on the theory of interval probability, is given in Section 2.1. In Section 2.2 an overview of nonparametric predictive inference (NPI) is presented and some fundamental properties of the NPI model for multinomial data are described. Section 2.3 contains a brief overview of two classification methods, namely naive classification and classification trees, and an explanation of how these methods have been adapted for use with interval probability models. An introduction to Weka software is also given, which is widely used in practical applications of classification methods. This software is used in Chapter 4.

2.1 Imprecise probability

In classical probability theory, the probability for an event A is given by a precise value $p(A) \in [0, 1]$, where p is a probability satisfying Kolmogorov's axioms. However, when information or knowledge is incomplete, a unique probability may be too restrictive. An alternative approach is to use imprecise probability, which is an umbrella term encompassing all qualitative and quantitative ways of measuring uncertainty without single-valued probabilities.

Imprecise probability is a long-standing concept (see Hampel [28] for a historical overview) and was first formally proposed in 1854 by Boole [12]. Throughout the twentieth century, several new theories were proposed for quantifying uncertainty, and the past two decades in particular have seen much new research in the field of imprecise probability including the development of interval probability theory. An interval probability consists of a lower probability $\underline{P} \in [0,1]$ and an upper probability $\overline{P} \in [0,1]$. The classical situation is a special case of interval probability with $\underline{P} = \overline{P}$ and the vacuous assessment $\underline{P} = 0$ and $\overline{P} = 1$ represents a total lack of knowledge about an event. Comprehensive foundations of interval probability theory have been proposed by Walley [40] and by Weichselberger [43, 44]. Walley's [40] treatment of the subject is based on De Finetti's [23] interpretation of probabilities as fair prices for gambles. According to Walley, the lower probability $\underline{P}(A)$ for some event A is interpreted as the maximum price for which one would buy the bet which pays 1 if A occurs and 0 otherwise. The upper probability $\overline{P}(A)$ is interpreted as the minimum price for which one would sell this bet. The theory developed by Walley [40] is based on a set of coherence conditions which ensures that the lower and upper probabilities are rational from the behavioural point of view. The foundations of interval probability presented by Weichselberger [43, 44] are purely theoretical with no assumed interpretation, generalising Kolmogorov's axioms. The theory and terminology developed by Weichselberger are used throughout this thesis and the relevant aspects of his approach to interval probability are explained below.

For a sample space Ω, the set of events $\mathcal{A}$ is given by the power set of Ω. An interval probability model assigns an interval probability $P(A) = [\underline{P}(A), \overline{P}(A)]$ to every event $A \in \mathcal{A}$, such that $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$ for all $A \in \mathcal{A}$. The structure of the model is defined by Weichselberger [43] as the set

$$\mathcal{M} = \{p | \underline{P}(A) \leq p(A) \leq \overline{P}(A), \forall A \in \mathcal{A}\}, \tag{2.1}$$

i.e. the set of all classical probability distributions p that are in accordance with the interval limits.

An interval probability is described by Weichselberger [43] as an F-probability if

$$\inf_{p\in\mathcal{M}} p(A) = \underline{P}(A)$$

and

$$\sup_{p\in\mathcal{M}} p(A) = \overline{P}(A)$$

for all $A \in \mathcal{A}$. For a finite set of events $\mathcal{A}$, the concept of F-probability coincides with Walley's [40] notion of coherence (see Weichselberger [43]). For a given F-probability, a prestructure is defined by Weichselberger [43] as any subset $\mathcal{J}$ of $\mathcal{M}$ such that $\inf_{p\in\mathcal{J}} p(A) = \underline{P}(A)$ for all $A \in \mathcal{A}$.

F-probability is an important concept in Weichselberger's theory, as it implies a number of other properties. When we have F-probability, lower probability is superadditive and upper probability is subadditive. This means that for events A and B, $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$ and $\overline{P}(A \cup B) \leq \overline{P}(A) + \overline{P}(B)$. Also, for every F-probability, the lower and upper probabilities satisfy the conjugacy relation $\overline{P}(A) = 1 - \underline{P}(A^c)$, where A^c is the complementary event to A. It can be shown [44] that on a finite space, $\mathcal{M}$ is a closed and convex set of probability distributions. Such sets are also called credal sets in the literature. The theory of credal sets of probability distributions has strong foundations (see Levi [31]) and is widely used [1, 40, 46–48].

In this thesis, we focus on interval probabilistic statistical inference, specifically for multinomial data. In a multinomial setting, observed data consist of n categorical observations $Y_1, ..., Y_n$. This data set comprises n_j observations in category c_j, for $j = 1, ..., K$. We consider inferences about a future observation Y_{n+1}.

One imprecise probability model for inference from multinomial data, Walley's Imprecise Dirichlet Model (IDM) [41], has attracted much attention and has been applied in a variety of areas [11]. The model generalises the standard Bayesian approach to inference from multinomial data. The multinomial model is assumed, where each category c_j is associated with a probability θ_j such that $\theta_j \geq 0$ and

$\sum_{j=1}^{K} \theta_j = 1$, and the vector $\underline{n} = (n_1, ..., n_K)$ follows a multinomial distribution with parameters n and $\underline{\theta}$ where $\underline{\theta}$ is the vector of probabilities $(\theta_1, ..., \theta_K)$. Under the IDM, prior uncertainty about $\underline{\theta}$ is described by the set of all Dirichlet $(s, \underline{t})$ distributions such that $0 < t_j < 1$ for $j = 1, ..., K$ and $\sum_{j=1}^{K} t_j = 1$. The choice of the parameter s is important and is discussed in detail by Walley [41]. After observing the data, Bayes' theorem is used to update each prior distribution in the set, and hence the posterior uncertainty about $\underline{\theta}$ is described by the set of all Dirichlet $(s, \underline{t^*})$ distributions where $t_j^* = \frac{n_j + st_j}{n+s}$ for $j = 1, ..., K$. Inferences about an event are in the form of a lower probability and an upper probability. These are obtained by minimising and maximising the posterior probability for the event with respect to $\underline{t}$. According to the IDM, the lower and upper probabilities for the event that the next observation will be in category c_j are $\underline{P}_{IDM}(Y_{n+1} \in c_j) = \frac{n_j}{n+s}$ and $\overline{P}_{IDM}(Y_{n+1} \in c_j) = \frac{n_j+s}{n+s}$. In the discussion of Walley's paper on the IDM [41] and indeed by Walley himself, a number of criticisms of the model were made. In the light of this, a new model for inference from multinomial data was proposed by Coolen and Augustin [14, 17]. This is an attractive alternative to the IDM which uses the theory of nonparametric predictive inference (NPI).

2.2 Nonparametric predictive inference

NPI is an inferential framework of statistical theory and methodology that is based on Hill's exchangeability-related assumption $A_{(n)}$ [29], defined below.

Definition 2.2.1. *Consider a real-valued data set consisting of observed values $X_i = x_i$, $i = 1, ..., n$, where we assume that there are no tied observations. These observations are ordered such that $x_1 < ... < x_n$ and they partition the real line into $n + 1$ open intervals (x_i, x_{i+1}) for $i = 0, ..., n$, where $x_0 = -\infty$ and $x_{n+1} = \infty$. The assumption $A_{(n)}$ states that the next observation will fall in any interval (x_i, x_{i+1}) with probability $\frac{1}{n+1}$, i.e. $P(X_{n+1} \in (x_i, x_{i+1})) = \frac{1}{n+1}$ for $i = 0, ..., n$.*

NPI makes use of $A_{(n)}$ to give predictive inferences about future observations in the form of lower and upper probabilities and has been presented for Bernoulli

data [13], real-valued data [6], data including right-censored observations [22] and multinomial data [14,17]. It has a wide range of applications in statistics, reliability and operational research, summarised by Coolen [16]. We focus on the NPI model for multinomial data [14, 17], which we henceforth refer to as the MNPI model. The MNPI model is based on a variation of Hill's assumption $A_{(n)}$ [29] called circular-$A_{(n)}$ that was first introduced in [14]. This assumption relates to circular data. Circular-$A_{(n)}$ is defined below.

Definition 2.2.2. *Consider a circular data set consisting of observed values* $Y_i = y_i$, $i = 1, ..., n$. *These observations create* n *intervals on a circle, which are represented as* $I_j = (y_j, y_{j+1})$ *for* $j = 1, ..., n-1$ *and* $I_n = (y_n, y_1)$. *The assumption circular-*$A_{(n)}$ *states that the next observation will fall in any interval* I_j *with equal probability* $\frac{1}{n}$, *i.e.* $P(Y_{n+1} \in I_j) = \frac{1}{n}$ *for* $j = 1, ..., n$.

The MNPI model involves a probability wheel representation of the data. On the probability wheel, each of the n categorical observations is represented by a radial line, such that the wheel is partitioned into n equally-sized slices. Making inferences about the next observation is analogous to spinning this probability wheel and from the circular-$A_{(n)}$ assumption we conclude that the next observation has probability $\frac{1}{n}$ of being in any given slice. The inferences given about a future observation therefore depend upon which category each slice of the wheel represents. The following assumption, and the constraints on the configuration of the wheel which this assumption implies, are key to the results presented in this thesis. Coolen and Augustin [14] assume that each category is only allowed to be represented by a single segment of the wheel, where a segment is defined as a single part of the wheel (note that the wheel is always divided radially) consisting of any number of full or partial slices. This assumption implies the following:

- Two or more lines representing observations in the same category must always be positioned next to each other on the wheel.

- A slice that is bordered by two lines representing observations in the same category must be assigned to that category. (An exception to this rule occurs when every line on the wheel represents the same category, in which case

$n-1$ slices must be assigned to that category and the remaining slice may be assigned to another category.)

- A slice that is bordered by two lines representing observations in categories c_i and c_j where $c_i \neq c_j$, defined as a separating slice, may be assigned to c_i or to c_j or to an unobserved category not yet allocated to any other slice.
- Separating slices may be divided radially between multiple categories.

The concept of the probability wheel is clarified further by the illustrative examples in the remainder of this section.

Throughout this work, it is assumed that there are K different categories altogether, and that k categories have already been observed and are labelled $c_1, ..., c_k$. It is assumed that $K \geq 3$: when $K = 2$, the MNPI model can be used, but the representation of the data on a line as presented for Bernoulli data [13] is more appropriate. Suppose that there are n_j observations in category c_j, for $j = 1, ..., k$. Therefore, $\sum_{j=1}^{k} n_j = n$. In some situations, we may know the total number K of possible categories, but at other times we may be unaware of this value. Coolen and Augustin presented lower and upper probabilities for the case where K is known [17] and for the case where K is unknown [14]. These results are summarised in Subsections 2.2.1 and 2.2.2.

2.2.1 Known number of categories

We summarise the results of Coolen and Augustin [17]. When K is known, events of interest can be expressed generally as

$$Y_{n+1} \in \bigcup_{j \in J} c_j \tag{2.2}$$

where $J \subseteq \{1, ..., K\}$. We refer to this general event as E. Let

$$OJ = J \cap \{1, ..., k\}$$

represent the index-set for the categories in E that have already been observed and let $r = |OJ|$. Also, let

$$UJ = J \cap \{k+1, ..., K\}$$

represent the index-set for the categories in E that have not yet been observed and let $l = |UJ|$. The derivation of the NPI lower and upper probabilities for the general event E is explained below.

Lower probability

We first consider the NPI lower probability for E. In order to find the minimal predictive probability, it is necessary to construct a configuration of the probability wheel which minimises the number of slices that must be assigned to E. This is done by separating lines representing different categories in E by categories not in E wherever possible.

We have r observed categories on the wheel which belong to E and we have $K - r - l$ categories which are not in E. We distinguish between the case where $r \leq K - r - l$ and the case where $r > K - r - l$, because when $r \leq K - r - l$ there are more categories not in E than there are observed categories in E. We consider these two cases separately. Note that in examples, E denotes specific events of the form shown in (2.2).

Case 1: $\boldsymbol{r \leq K - r - l}$

When the number of possible categories not in E is greater than or equal to the number of observed categories in E, all lines on the wheel representing different categories in E can be separated by categories not in E. Therefore, the only slices which we are forced to assign to E are those between two lines representing the same category in E. This leads to

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n}. \tag{2.3}$$

Example 2.2.1. Consider a multinomial data set with possible categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). The data are

$$(n_B, n_G, n_R, n_Y, n_P, n_O) = (4, 2, 1, 1, 0, 0).$$

Suppose that we are interested in the event $Y_9 \in \{\text{B,G,P}\}$. Then $K = 6$, $r = 2$ and $l = 1$, so this example illustrates the situation where $r \leq K - r - l$. We can therefore find a configuration of the probability wheel such that all categories in E are separated by categories not in E. Figure 2.1 shows one such configuration,

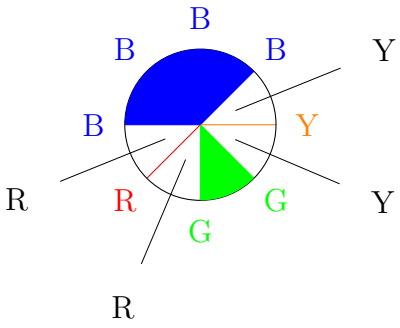

Figure 2.1: Probability wheel for Example 2.2.1

where B and G are separated by R and Y. The only slices of the wheel which must be assigned to E are the slices between two B observations or two G observations. This means that the NPI lower probability for E is $\frac{4}{8}$. This is verified by (2.3). $\diamond$

Case 2: $r > K - r - l$

When the number of observed categories in E is greater than the number of categories not in E, we cannot separate all lines on the wheel representing different categories in E. We separate as many of these as possible, but there are $r - (K - r - l) = 2r + l - K$ separating slices that cannot be assigned to a category not in E. These slices must therefore be assigned to E, leading to

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{2r + l - K}{n}. \tag{2.4}$$

Example 2.2.2. Consider the data set described in Example 2.2.1. Suppose that we are interested in the event $Y_9 \in \{\text{B,G,P,R,Y}\}$. Then $K = 6$, $r = 4$ and $l = 1$, so we

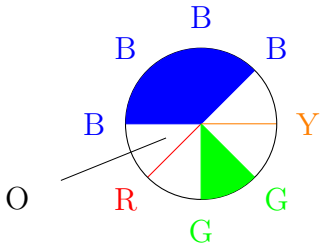

Figure 2.2: Probability wheel for Example 2.2.2

have the situation where $r > K - r - l$. We therefore cannot find any configuration of the probability wheel such that all observed categories in E are separated by categories not in E. Figure 2.2 shows a configuration of the probability wheel corresponding to the NPI lower probability. As before, the slices between two B observations or two G observations must be assigned to B and G respectively. We can assign one of the remaining slices to O, but there are no other categories not in E and therefore the remaining three separating slices must also be assigned to E. This means that the NPI lower probability for E is $\frac{7}{8}$. This is verified by (2.4). $\diamond$

Upper probability

We now consider the NPI upper probability for the general event E (2.2). In terms of the probability wheel, we want to assign as many slices as possible to E. The best configuration will therefore involve separating observed categories not in E by categories in E wherever possible.

We have $k - r$ observed categories on the wheel which are not in E and we have $r + l$ categories altogether which are in E. We consider separately the case $k - r \leq r + l$ and the case $k - r > r + l$.

Case 1: $k - r \leq r + l$

When the number of categories in E is not less than the number of observed categories not in E, all lines on the wheel representing different categories not in E can be separated by categories in E.

Therefore, all k separating slices can be assigned to E. Furthermore, we must assign to E all slices between two lines representing the same category in E. This leads to

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{k}{n}. \tag{2.5}$$

Example 2.2.3. Consider the data set described in Example 2.2.1 and suppose that we are interested in the event $Y_9 \in \{\text{B,G,P}\}$. Then $k = 4$, $r = 2$ and $l = 1$, so this example illustrates the situation where $k - r \leq r + l$. We can therefore find a configuration of the probability wheel such that all categories not in E are separated by categories in E. Figure 2.3 shows one possible configuration of the probability

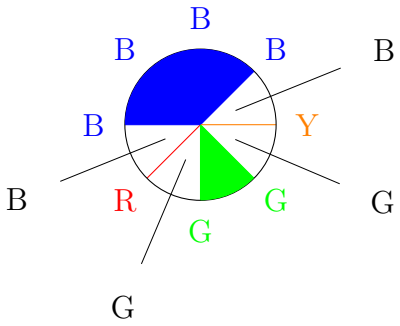

Figure 2.3: Probability wheel for Example 2.2.3

wheel, where the slices separating B from R and Y are both assigned to B and the slices separating G from R and Y are both assigned to G. The slices between two B observations or two G observations are also assigned to B and G respectively. This means that the NPI upper probability for E is 1. This is verified by (2.5). ◇

Case 2: $k - r > r + l$

When the number of observed categories not in E is greater than the number of categories in E, there is no configuration of the wheel such that all categories not in E are separated by categories in E. This means that we cannot assign all k separating slices to E. There are $k - r - (r + l) = k - 2r - l$ separating slices that cannot be assigned to E and therefore only $k - (k - 2r - l) = 2r + l$ separating slices can be assigned to E. We must also assign to E all slices between lines representing

the same category in E. This leads to

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{2r + l}{n}. \tag{2.6}$$

Example 2.2.4. Consider again the data set described in Example 2.2.1 and suppose that we are interested in the event $Y_9 \in \{\text{B,P}\}$. Then $k = 4$, $r = 1$ and $l = 1$, so we have the situation where $k - r > r + l$. We therefore cannot find any configuration of the probability wheel such that all observed categories not in E are separated by categories in E. Figure 2.4 shows a configuration of the probability

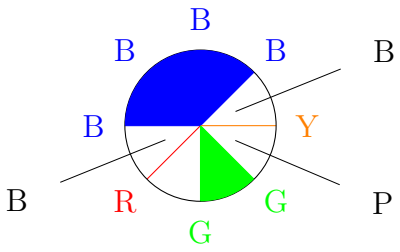

Figure 2.4: Probability wheel for Example 2.2.4

wheel corresponding to the NPI upper probability, where the slices separating B from R and Y are both assigned to B and the slice separating G from Y is assigned to P. The slices between two B observations are assigned to B and the slice between the two G observations is assigned to G. This means that the NPI upper probability for E is $\frac{6}{8}$. This is verified by (2.6). ◇

2.2.2 Unknown number of categories

We summarise the results of Coolen and Augustin [14]. When K is unknown, an event can no longer be expressed using the subset $J \subseteq \{1, ..., K\}$, so new notation is introduced. As before, $c_1, ..., c_k$ are the observed categories. In addition we define DN_i, $i = 1, ..., l$, as Defined New categories, which represent categories we wish to specify in the event of interest that have not yet been observed. We also define UN as the set of Unobserved New categories, which refers to any not yet observed category. Note that the categories DN_i are contained within the set UN. The derivation of the NPI lower and upper probabilities is based on the assumption that there is no

finite limit on the number of UN categories. Coolen and Augustin [14] give two general expressions which encompass any event of interest. These expressions are

$$Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \setminus \bigcup_{i=1}^{l} DN_i \tag{2.7}$$

and

$$Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup \bigcup_{i=1}^{l} DN_i \tag{2.8}$$

where $c_{j_1}, ..., c_{j_r}$ are the observed categories in the event of interest, $l \geq 0$ and $0 \leq r \leq k$. We refer to these general events as E_1 (2.7) and E_2 (2.8). The derivation of the NPI lower and upper probabilities for each of these general events is explained below.

Lower probability

We consider the NPI lower probabilities for events E_1 and E_2. As in the situation where K is known, we find the NPI lower probability by constructing a configuration of the probability wheel which separates lines representing different categories in the event of interest by categories not in the event of interest. This minimises the number of slices that must be assigned to the event of interest.

We first consider E_1. There are k observed categories and therefore there are k separating slices. If $k \geq 2r$, at least half of the k observed categories are not in E_1. We can therefore find a configuration of the wheel which separates all observed categories in E_1 by observed categories not in E_1, meaning that the only slices which we are forced to assign to E_1 are those between two lines representing the same category in E_1. However, if $r \leq k < 2r$, not all observed categories in E_1 can be separated. There are $r - (k - r) = 2r - k$ separating slices that cannot be assigned to an observed category not in E_1. We can, however, assign l of these slices to categories DN_i, $i = 1, ..., l$, since these are not in E_1. Overall, there are

$\max(2r-k-l,0)$ slices which we are forced to assign to E_1. This leads to

$$\underline{P}(E_1) = \begin{cases} \displaystyle\sum_{s=1}^{r} \frac{n_{j_s}-1}{n} & \text{if } k \geq 2r \\ \displaystyle\sum_{s=1}^{r} \frac{n_{j_s}-1}{n} + \frac{\max(2r-k-l,0)}{n} & \text{if } r \leq k < 2r \end{cases} \tag{2.9}$$

where n_{j_s} is the number of times c_{j_s} has been observed.

Example 2.2.5. Consider a multinomial data set where the set of possible categories consists of an unknown number of different colours. We have observed the following categories: blue (B), green (G), red (R) and yellow (Y). We define two new categories: pink (P) and orange (O). The data are

$$(n_B, n_G, n_R, n_Y) = (4, 2, 1, 1).$$

Suppose that we are interested in the event $Y_9 \in \{(\mathrm{B} \cup \mathrm{G}) \cup UN \setminus (\mathrm{P} \cup \mathrm{O})\}$. Then $k = 4$, $r = 2$ and $l = 2$. This event is of type E_1 and this example illustrates the situation where $k \geq 2r$. We can therefore configure the probability wheel such that all observed categories in the event of interest are separated by observed categories not in the event of interest. Figure 2.5 shows one such configuration of the wheel,

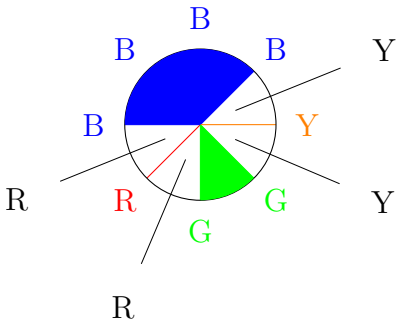

Figure 2.5: Probability wheel for Example 2.2.5

where B and G are separated by R and Y. The only slices of the wheel which must be assigned to the event of interest are the slices between two B observations or two G observations. This means that the NPI lower probability for the event of interest is $\frac{4}{8}$. This is verified by (2.9). ◇

Example 2.2.6. Consider again the data set described in Example 2.2.5. Suppose that we are interested in the event $Y_9 \in \{(\mathrm{B} \cup \mathrm{G} \cup \mathrm{R}) \cup UN \setminus \mathrm{P}\}$. Then $k = 4$, $r = 3$ and $l = 1$. This event is also of type E_1, but this example illustrates the situation

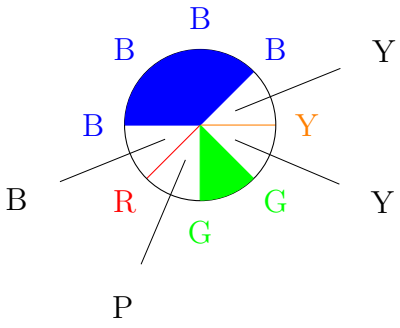

Figure 2.6: Probability wheel for Example 2.2.6

where $k < 2r$. We cannot find a configuration of the probability wheel such that all observed categories in the event of interest are separated by observed categories not in the event of interest. However, we can assign one separating slice to the Defined New category P that is excluded from the event of interest. Figure 2.6 shows such a configuration, where B and G are separated by Y, and also P separates G and R. The remaining separating slice has to be assigned to the event of interest and is assigned to B in Figure 2.6. The slices between two B observations or two G observations must also be assigned to the event of interest. This means that the NPI lower probability for the event of interest is $\frac{5}{8}$. This is verified by (2.9). $\diamond$

We now consider E_2. As before, we have k observed categories, so we have k separating slices. The event E_2 only includes l of the UN categories, so since no maximum number of categories in UN is assumed, we can suppose that all of the k separating slices can be assigned to a category not in E_2, be it observed or unobserved. Therefore, the only slices which must be assigned to E_2 are those between two lines representing the same category in E_2. This leads to

$$\underline{P}(E_2) = \sum_{s=1}^{r} \frac{n_{j_s} - 1}{n}. \tag{2.10}$$

Example 2.2.7. Consider the data set described in Example 2.2.5. Suppose that we are interested in the event $Y_9 \in \{(\mathrm{B} \cup \mathrm{G} \cup \mathrm{R}) \cup (\mathrm{P} \cup \mathrm{O})\}$. Then $k = 4$, $r = 3$ and

$l = 2$. This event is of type E_2. We can configure the probability wheel such that all

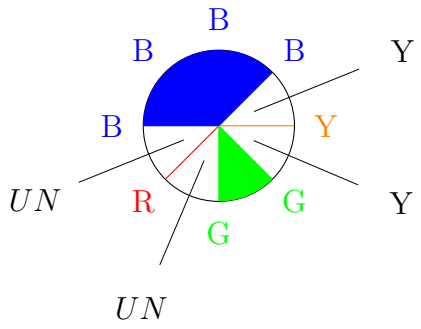

Figure 2.7: Probability wheel for Example 2.2.7

categories in the event of interest are separated either by observed categories not in the event of interest or by UN categories. Figure 2.7 shows one such configuration of the wheel. Here, B and G are separated by Y. The event of interest does not include any UN categories except P and O, so we use UN categories other than P and O to separate B and R and to separate R and G. (Note that the two slices assigned to UN must represent different Unobserved New categories.) Therefore, the only slices which we must assign to the event of interest are the slices between two B observations and the slices between two G observations. This means that the NPI lower probability for the event of interest is $\frac{4}{8}$. This is verified by (2.10). $\diamond$

Upper probability

We consider the NPI upper probabilities for events E_1 (2.7) and E_2 (2.8). Here, we wish to find a configuration of the probability wheel which assigns as many slices as possible to the event of interest.

We first consider E_1. As described previously, there are k separating slices on the wheel. Since E_1 contains all except a finite number of the UN categories, we can assume that every one of these k slices can be assigned either to an observed category in E_1 or to an unobserved category in E_1. Furthermore, we assign to E_1 any slices between two lines representing the same category in E_1. This leads to

$$\overline{P}(E_1) = \sum_{s=1}^{r} \frac{n_{j_s} - 1}{n} + \frac{k}{n}. \tag{2.11}$$

Example 2.2.8. Consider the data set described in Example 2.2.5. Suppose that we are interested in the event $Y_9 \in \{\mathrm{B} \cup UN \setminus (\mathrm{P} \cup \mathrm{O})\}$. Then $k = 4$, $r = 1$ and $l = 2$. This event is of type E_1 and we can configure the probability wheel such that all

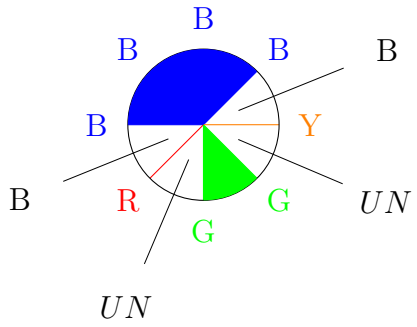

Figure 2.8: Probability wheel for Example 2.2.8

categories not in the event of interest are separated either by observed categories in the event of interest or by unobserved categories in the event of interest. Figure 2.8 shows one such configuration of the wheel. Here, R and Y are separated by B. The event of interest includes all UN categories except P and O, so we use UN categories other than P and O to separate R and G and to separate G and Y. This means that the NPI upper probability for the event of interest is $\frac{7}{8}$. This is verified by (2.11). ◇

We now consider E_2. Again, we consider the k separating slices on the probability wheel. If $k \leq 2r$, at least half of the k observed categories are in E_2, so we can construct a configuration of the wheel where all observed categories not in E_2 are separated by observed categories in E_2. Therefore, all k separating slices can be assigned to E_2. Conversely, if $k > 2r$, we cannot separate all observed categories not in E_2. Two of the separating slices can be assigned to each observed category in E_2, but we will have $k - 2r$ remaining separating slices that cannot be filled by an observed category in E_2. However, E_2 also contains l unobserved categories, namely DN_i, $i = 1, ..., l$. This means that we can assign up to l of the separating slices to these categories. Overall, there will be $\min(k - 2r, l)$ remaining separating slices which can be assigned to E_2. We also assign to E_2 all slices between lines representing the same category in E_2. This leads to

$$\overline{P}(E_2) = \begin{cases} \displaystyle\sum_{s=1}^{r} \frac{n_{j_s} - 1}{n} + \frac{k}{n} & \text{if } r \leq k \leq 2r \\ \displaystyle\sum_{s=1}^{r} \frac{n_{j_s} + 1}{n} + \frac{\min(k - 2r, l)}{n} & \text{if } k > 2r \end{cases} \tag{2.12}$$

Example 2.2.9. Consider the data set described in Example 2.2.5. Suppose that we are interested in the event $Y_9 \in \{(\mathrm{B} \cup \mathrm{G}) \cup (\mathrm{P} \cup \mathrm{O})\}$. Then $k = 4$, $r = 2$ and $l = 2$. This event is of type E_2, and this example illustrates a situation where $k \leq 2r$. We can therefore configure the probability wheel such that all observed categories not in the event of interest are separated by observed categories in the event of interest. Figure 2.9 shows one such configuration of the wheel, where R and Y are separated

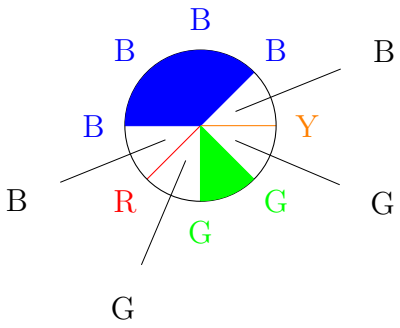

Figure 2.9: Probability wheel for Example 2.2.9

by B and G. This means that the NPI upper probability for the above event is 1. We can verify this using (2.12). ◇

Example 2.2.10. Consider once more the data set described in Example 2.2.5. Suppose that we are interested in the event $Y_9 \in \{\mathrm{B} \cup (\mathrm{P} \cup \mathrm{O})\}$. Then $k = 4$, $r = 1$ and $l = 2$. This event is of type E_2 and we have the situation where $k > 2r$. This means we cannot separate all observed categories not in the event of interest by observed categories in the event of interest. However, we have two Defined New categories, P and O, which are included in the event of interest. Figure 2.10 shows one configuration of the wheel corresponding to upper probability, where R and Y are separated by B, and also P separates R and G, and O separates G and Y. This

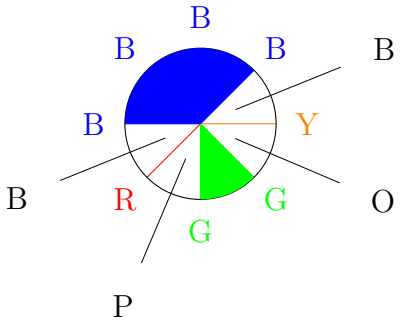

Figure 2.10: Probability wheel for Example 2.2.10

means that the NPI upper probability for the event of interest is $\frac{7}{8}$. This is verified by (2.12). ◇

2.2.3 Properties of the model

It was proven by Coolen and Augustin [14, 17] that the NPI lower and upper probabilities satisfy the conjugacy property $\overline{P}(E) = 1 - \underline{P}(E^c)$, where E^c represents the complementary event to E, i.e. an event containing all possible categories except for those contained in E. It can also be proven that $\underline{P}(E) \leq \sum_{c_j \in E} \frac{n_j}{n} \leq \overline{P}(E)$, showing that the model is a generalisation of the naive model in which the probability for a category is equal to the relative frequency of that category in the data. It can be shown that the imprecision vanishes as $n \to \infty$. Coolen and Augustin [14, 17] also proved that the intervals given by the NPI lower and upper probabilities are F-probabilities in the sense of Weichselberger [43]. This implies coherence in Walley's sense [40] at any single point in time, meaning that the NPI lower and upper probabilities are rational from the behavioural point of view. The issues of conditioning and updating are not explored in this thesis, but it is important to note the difference between these actions: in the NPI framework, updating is not done in a Bayesian manner through conditioning, but is done by taking new data into account as well as the previously available data and recalculating the NPI lower and upper probabilities based on the new total number of observations. It can be shown that the NPI approach has strong consistency properties for both conditioning and updating and this is discussed

further in [6], [17] and [18].

The set of probability distributions that are consistent with the MNPI model is defined by the constraints of the probability wheel, explained at the beginning of this section. Any distribution belonging to this set must clearly belong to the structure $\mathcal{M}$ of the model. However, not all distributions in $\mathcal{M}$ correspond to a valid configuration of the probability wheel. This is discussed further in Chapter 4. The set of probability distributions that are in accordance with the MNPI model is henceforth referred to as the NPI structure, and is denoted by $\mathcal{M}_{NPI}$, where $\mathcal{M}_{NPI} \subseteq \mathcal{M}$. We see that $\mathcal{M}_{NPI}$ is a prestructure in the sense of Weichselberger [43].

2.3 Classification

Classification is a procedure in which units are categorised according to the observed values of one or more attribute variables. Suppose that $\{X_i\}_{i=1}^m$ is the set of attribute variables and that X_i has possible values $x_i = 1, ..., |A_i|$ in a finite set A_i, where $|A_i|$ is the number of elements in A_i. Then a classifier maps a set of attribute values $\{x_1, ..., x_m\}$ to a category $C \in \{c_1, ..., c_K\}$. Classifiers are learned from a training set of data, which is a sample of n observations for which the category is already known. Recent advances in imprecise probability theory have led to the development of classification methods based on imprecise probabilities, the most widely used of which are naive classification and classification trees.

2.3.1 Naive classification

The naive Bayes classifier, proposed by Duda and Hart [25], is a straightforward method based on the posterior probabilities $p(C|X_1, ..., X_m)$, $j = 1, ..., K$. The naive assumption is made that each attribute is conditionally independent of every other attribute given C, so the application of Bayes' theorem gives $p(C|X_1, ..., X_m) \propto p(C) \prod_{i=1}^{m} p(X_i|C)$. The probabilities $p(C)$ and $p(X_i|C)$ are estimated from the training set. Given a new observation with attribute

values $x_1, ..., x_m$, the output from the naive Bayes classifier is given by $\underset{c_j}{\text{argmax}}\ p(C = c_j)\prod_{i=1}^{m} p(X_i = x_i|C = c_j)$, i.e. the category which has the largest posterior probability.

The naive credal classifier, proposed by Zaffalon [46, 48], is an extension of the naive Bayes classifier to credal sets of probability distributions. The set $\mathcal{P}_C$ is defined to be a credal set of distributions $p(C)$ and for each $j \in \{1, ..., K\}$ the set $\mathcal{P}_{X_i}^{c_j}$ is defined to be a credal set of distributions $p(X_i|c_j)$. The naive assumption together with Bayes' theorem then gives the set of joint distributions $\mathcal{P} = \{p(C)\prod_{i=1}^{m} p(X_i|C)|p(C) \in \mathcal{P}_C, p(X_i|c_j) \in \mathcal{P}_{X_i}^{c_j}, i = 1, ..., m, j = 1, ..., K\}$. The credal sets $\mathcal{P}_C$ and $\mathcal{P}_{X_i}^{c_j}$ may be obtained using statistical inference, as discussed in Section 2.1, or they may be provided by subjective judgements. Given a new observation with attribute values $x_1, ..., x_m$, lower and upper probabilities for each category c_j are obtained by minimising and maximising the posterior probability $p(c_j|x_1, ..., x_m)$ with respect to the set $\mathcal{P}$. This optimisation problem is discussed in detail by Zaffalon [46, 48]. The use of credal sets of distributions means that the output of the naive credal classifier is generally a set of categories instead of a single category, because the interval probabilities for categories may overlap. This is referred to as imprecise classification. Various methods are proposed for comparing these interval probabilities [46, 48].

Zaffalon [47] showed that the IDM [40] can be used to infer the credal sets that define the naive credal classifier. However, due to the parametric nature of this method of classification, the MNPI model is not a natural alternative to the IDM for this particular application. We therefore do not further consider naive classification in this work.

2.3.2 Classification trees

A classification tree is a hierarchical structure used for predicting the category of an observation from the values of the attribute variables. Within a classification tree,

each node represents a single attribute variable, referred to as the split variable for the node, and each possible value of this variable corresponds to a particular branch from the node. We restrict attention to discrete attribute variables here, so the number of splits at a node is defined by the number of possible values of the split variable. Each leaf of the tree represents a category, and when classifying an observation, the combination of attribute values observed defines a path through the tree from the root node to a leaf node, leading to a prediction about the category of this observation. Example 2.3.1 illustrates this method of classification.

Example 2.3.1. Consider a data set with two possible categories, $C1$ and $C2$, and two attribute variables, X and Y. X has possible values x_1 to x_4 and Y has possible values y_1 and y_2. Suppose that we wish to classify an observation which has attribute

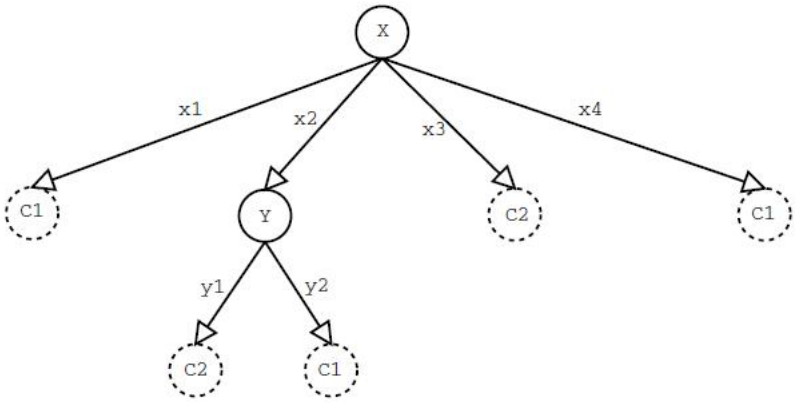

Figure 2.11: Classification tree for Example 2.3.1

values $X = x_2$ and $Y = y_1$. Using the classification tree shown in Figure 2.11, we see that the predicted category for this observation is $C2$. ◇

Classification trees are constructed using a subset of the observed data called the training set. The remaining data, called the test set, are then used to test the performance of the classifier. In practical applications, a ten-fold cross-validation procedure is commonly used [30]. The observed data are randomly partitioned into ten subsets, of which nine subsets are used as the training set and the remaining subset is used as the test set. The tree-building process is repeated ten times, and

each of the ten subsets is used precisely once as the test set. The ten results are then averaged.

When building a classification tree using the training set, an impurity measure is used to select the split variable at each node. Such a measure quantifies the extent of the variation between the categories of the observations at that node and allows us to determine the information content of an attribute variable. At the root node, the set of all attribute variables is considered and a variable is selected such that the impurity reduction achieved by the split is maximised. We then proceed in a similar way for further nodes, splitting on the most informative remaining attribute variable at each stage. When no further impurity reduction is achievable, we stop splitting and a leaf node is produced showing the most probable category.

There are several different versions of this tree-building process that use various impurity measures [34, 35, 45], but we focus now on Quinlan's ID3 method [34]. This method is based on the principle of Occam's Razor, in the sense that smaller classification trees are preferred to larger ones, and this principle is applied through the use of an impurity measure based on the Shannon entropy [38]

$$H(p) = -\sum_{j=1}^{K} p(c_j) \log[p(c_j)], \tag{2.13}$$

which is a function of the true category probabilities $p(c_j)$. Since these probabilities are unknown, it is necessary to use the estimator

$$\widehat{H}(\widehat{p}) = -\sum_{j=1}^{K} \widehat{p}(c_j) \log[\widehat{p}(c_j)] \tag{2.14}$$

for the Shannon entropy, where $\widehat{p}(c_j)$ is an estimate of $p(c_j)$. Suppose that at some node R we have n^R observations in total, of which n_j^R belong to category c_j for $j = 1, ..., K$. At this node, the Shannon entropy is estimated by

$$\widehat{H}(\widehat{p}^R) = -\sum_{j=1}^{K} \widehat{p}^R(c_j) \log[\widehat{p}^R(c_j)] \tag{2.15}$$

where $\widehat{p}^R(c_j)$ indicates the relative frequency $\frac{n_j^R}{n^R}$. The attribute variable X_i is then considered for splitting. At each new node produced by this split, we have $n^{R \cup (X_i = x_i)}$

observations in total, of which $n_j^{R\cup(X_i=x_i)}$ belong to category c_j for $j = 1, ...K$. At this node, the Shannon entropy is estimated by

$$\widehat{H}(\widehat{p}^{R\cup(X_i=x_i)}) = -\sum_{j=1}^{K} \widehat{p}^{R\cup(X_i=x_i)}(c_j) \log[\widehat{p}^{R\cup(X_i=x_i)}(c_j)] \tag{2.16}$$

where $\widehat{p}^{R\cup(X_i=x_i)}(c_j)$ indicates the relative frequency $\frac{n_j^{R\cup(X_i=x_i)}}{n^{R\cup(X_i=x_i)}}$. The weighted sum of impurity measures over all new nodes is

$$I(R, X_i) = \sum_{x_i \in A_i} \frac{n^{R\cup(X_i=x_i)}}{n^R} \widehat{H}(\widehat{p}^{R\cup(X_i=x_i)}), \tag{2.17}$$

and the impurity reduction achieved by splitting on X_i is therefore equal to $\widehat{H}(\widehat{p}^R) - I(R, X_i)$. However, since $\widehat{H}(\widehat{p}^R)$ is constant for all X_i, the impurity reduction is measured by $I(R, X_i)$ for practical purposes.

The ID3 method [34] can be adapted for use with interval probability models, as shown by Abellan and Moral [1]. At each node of the tree, an interval probability model is used to give $P(c_j) = [\underline{P}(c_j), \overline{P}(c_j)]$ for each category. The impurity measure $\widehat{H}$ (2.14) based on the Shannon entropy is still used, but the true category probabilities are now estimated using the maximum entropy distribution p_{maxE}, which is the distribution that maximises (2.14) taken from the set of all possible probability distributions that are consistent with the interval probability model. The impurity reduction achieved by splitting on X_i is measured by

$$I(R, X_i) = \sum_{x_i \in A_i} \frac{n^{R\cup(X_i=x_i)}}{n^R} \widehat{H}(p_{maxE}^{R\cup(X_i=x_i)}). \tag{2.18}$$

As we are using interval probabilities, it may not always be clear which is the most probable category at a leaf node. As mentioned in Subsection 2.3.1, there are various ways of comparing the interval probabilities $P(c_j)$ [46, 48]. However, for our purposes, we restrict to the method whereby if we cannot give a single most probable category at a leaf node, the parent node is considered and the most probable category at that node is used.

Abellan and Moral [1] considered the use of the IDM [41] for building

classification trees. The maximum entropy distributions were taken from the credal set of distributions associated with the IDM lower and upper probabilities $\underline{P}_{IDM}(Y_{n+1} \in c_j)$ and $\overline{P}_{IDM}(Y_{n+1} \in c_j)$ (see Section 2.1). However, given the nonparametric nature of classification trees, the MNPI model would be a suitable alternative to the IDM for this application. This is considered further in Chapter 4.

2.3.3 Weka software

Weka is a suite of machine learning software that was developed at the University of Waikato in New Zealand. It contains implementations of many classification methods as well as a variety of tools, called filters, that are used for preprocessing data. It is also possible to implement new classification methods in Weka. There are two main graphical user interfaces to Weka: the Explorer and the Experimenter. The Explorer enables the user to apply a classification method to a single data set and to analyse the performance of the resulting classifier. The Experimenter allows the user to apply multiple classification methods to a collection of data sets and to compare the performance of the methods via statistical tests. The software is available on the Weka webpage (www.cs.waikato.ac.nz/ml/weka), where current information on the software can also be found [27]. A comprehensive guide to Weka is given in [45]. We make use of this software for the practical applications seen in Chapter 4.

Chapter 3

Selection

In this chapter we present methods based on nonparametric predictive inference (NPI) for selection problems involving multinomial data. Selection is a wide-ranging topic in statistics which involves finding methods for choosing the optimal member(s) of some group. This group may be, for example, a set of categories or a range of data sources. With regard to multinomial data, interest may be in choosing the category that has the largest probability of occurrence. Bechhofer et al. [10] present an overview of existing methods for this type of selection, but these are based on hypothesis testing and do not consider predictive inference. These methods are described briefly in Section 3.1. NPI-based selection has been applied to other problems, such as the selection of an optimal data source where observations are real-valued [21], for which the possibility of right-censored observations in the data has also been considered [32], and the selection of an optimal group of Bernoulli data [19, 20]. An overview of these types of predictive selection is given in Section 3.2. In the rest of the chapter, we discuss the use of NPI for category selection for a multinomial data set. We present a method for selection of a single category in Section 3.3 and a method for selection of a subset of categories in Section 3.4.

3.1 Overview of category selection methods

Given a set of multinomial data, we wish to select the category which is most likely to occur, or a subset of categories containing this category. Suppose that there are K

possible categories, $c_1, ..., c_K$, with associated unknown probabilities $p_1, ..., p_K$ where p_j is the classical probability that an observation is in category c_j. Let $n_1, ..., n_K$ denote the number of observations in categories c_1 to c_K, respectively. Suppose that the category probabilities are reordered in an increasing way and relabelled such that we have $p_{[1]} \leq ... \leq p_{[K]}$. The objective is therefore to select the category associated with $p_{[K]}$, or a subset of the K categories which contains this category. The following selection methods were summarised by Bechhofer et al. [10].

3.1.1 Indifference zone procedures

The indifference zone approach to selection was first introduced by Bechhofer [9]. The objective here is to select the single category associated with $p_{[K]}$, defined as correct selection (CS). It is required that $p(CS) \geq P^*$ whenever $\frac{p_{[K]}}{p_{[K-1]}} \geq \Theta^*$, for specified constants P^* and Θ^* such that $1 < \Theta^* < \infty$ and $\frac{1}{K} < P^*$. The constant Θ^* is defined as the indifference zone. Θ^* and P^* are specified prior to sampling and are used to determine the total number of observations, n, that should be sampled. The sample is then taken and category c_j is selected such that $n_j = \max_{i \in 1,...,K} n_i$. The use of curtailment is possible, where sampling can be stopped after $n' < n$ observations if $n_j \geq n_i + n - n'$ for all $i \neq j$.

3.1.2 Subset procedures

The subset selection approach was first introduced by Gupta [26]. The objective here is to find a subset of categories which contains the category associated with $p_{[K]}$. The selection of such a subset is defined as correct selection (CS). It is required that $p(CS) \geq P^*$ for a specified constant P^*. Selection is based on a sample of n observations and category c_j is included in the selected subset if and only if $n_j \geq \max_{i \in 1,...,K} n_i - d$ for a specified value d which depends on K, n and P^*.

3.2 Overview of predictive selection methods

Selection methods based on NPI have been developed by Coolen and van der Laan [21] and Coolen and Coolen-Schrijner [19, 20]. These methods use predictive

inferences based on past observations and make use of Hill's assumption $A_{(n)}$ [29].

Coolen and van der Laan [21] developed a NPI selection method for real-valued data from k different sources. Their objective was to select the source which will provide the largest next observation and this was done by making inferences about one future observation from each source. NPI lower and upper probabilities were determined for the event that the next observation from one source will exceed the next observation from all other sources. They also considered two ways of selecting a subset of sources: first, they determined the interval probability that some subset will contain the source providing the largest next observation and secondly, they found the interval probability that the next observations from every source in some subset will all exceed the next observations from the remaining sources.

Coolen and Coolen-Schrijner [19, 20] developed a NPI selection method for Bernoulli data from k different groups. The objective was to select the group which will have the highest number of future successes. Here, inferences were made about m future observations as opposed to a single observation. Groups were compared pairwise [20] to find the NPI lower and upper probabilities for the event that one group will have more future successes than another and a multiple comparison was carried out [20] to find the NPI lower and upper probabilities for the event that one group will have more future successes than any other group. Subsets of the groups were also considered [19], with interval probabilities presented for the event that some subset contains the group which has the most future successes and for the event that all groups in some subset will have more future successes than every other group.

3.3 NPI category selection for multinomial data

We develop NPI-based methods for category selection for a multinomial data set. We have K possible categories, labelled $c_1, ..., c_K$, where K is a known value, and our aim is to select a single category with the largest NPI lower or upper probability

for some event of interest involving future observations. Suppose that we have a data set consisting of n observations and let $n_1, ..., n_K$ denote the number of times we have observed categories $c_1, ..., c_K$, respectively. We consider m future observations and we select a category based on predictive inferences about these m observations. These inferences are made by using and extending the general theory of NPI [17], discussed in Chapter 2. For $j = 1, ..., K$, let M_j denote the random quantity representing the number of the m future observations that belong to category c_j, so $\sum_{j=1}^{K} M_j = m$.

3.3.1 One future observation

The simplest case is where $m = 1$, so inference is regarding one future observation. We consider the problem of selecting a single category with the largest NPI lower or upper probability for the event that the next observation belongs to this category. According to the MNPI model, the lower and upper probabilities for the event that the next observation will belong to category c_j are

$$\underline{P}(M_j = 1) = (\frac{n_j - 1}{n})^+, \tag{3.1}$$

where the notation x^+ is used to represent $\max\{x, 0\}$, and

$$\overline{P}(M_j = 1) = \min\{\frac{n_j + 1}{n}, 1\}. \tag{3.2}$$

We see that these lower and upper probabilities are monotonically increasing in n_j. We can evaluate these probabilities for each of the possible categories and then select the category with the largest NPI lower or upper probability.

Example 3.3.1. Consider a multinomial data set with possible categories blue (B), red (R), yellow (Y) and green (G). The data are

$$(n_B, n_G, n_R, n_Y) = (3, 2, 2, 1).$$

Suppose that we want to select a single category with the largest NPI lower and upper probability for the event that the next observation is in that category. First, we find the NPI lower and upper probability for the event that the next observation is

in B. By (3.1) and (3.2) we have $\underline{P}(M_B = 1) = \frac{n_j - 1}{n} = \frac{2}{8}$ and $\overline{P}(M_B = 1) = \frac{n_j + 1}{n} = \frac{4}{8}$. For the other categories, $P(M_Y = 1) = P(M_G = 1) = [\frac{1}{8}, \frac{3}{8}]$ and $P(M_R = 1) = [0, \frac{2}{8}]$, so we select B. ◇

Theorem 3.3.1. *When $m = 1$ and we want to select a single category with the largest NPI lower or upper probability for the event that the next observation belongs to that category, it is always optimal to choose the category which has the greatest number of observations in the data set.*

Proof. This follows directly from (3.1) and (3.2). These NPI lower and upper probabilities increase with n_j, so it is optimal to select the category with the largest value of n_j, i.e. the greatest number of observations. □

3.3.2 Multiple future observations

Whereas Coolen and Augustin [17] only considered one future observation, we now consider inferences about multiple future observations, so $m > 1$. Suppose that the data set is represented on a probability wheel and the n slices on the wheel are numbered 1 to n. Each of the m future observations must fall in one of these n slices. Let the vector $(S_1, ..., S_n)$ denote the number of future observations which fall in slices 1 to n respectively. The total number of different arrangements of these m observations is equal to $\binom{n+m-1}{m}$. Coolen [15] shows that it follows from the circular-$A_{(n)}$ assumption that each of these arrangements is equally likely, giving the precise probability for a particular arrangement

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

where $s_j \geq 0$ and $\sum_{j=1}^{n} s_j = m$.

More generally, the total number of different arrangements of f future observations within a segment made up of S slices, or $S + 1$ observations, is equal to

$$\binom{(S-1)+f}{f}. \tag{3.3}$$

This is because there are $S-1$ existing observations within the interior of such a segment, so we are considering the number of arrangements of f future observations amongst a total of $(S-1)+f$ observations.

(3.3) is needed when making inferences about multiple future observations. We first consider the case $m = 2$.

Two future observations

When $m = 2$, it may be of interest to consider the probability for the event that precisely one of the two future observations is in category c_j. In terms of the probability wheel, this means that one of the future observations falls in a slice allocated to category c_j and one falls elsewhere. Figure 3.1 illustrates the relevant segments of the wheel. We first consider the NPI lower probability for this event. There are $n_j - 1$ slices of the wheel which must be assigned to observed category c_j, corresponding to the shaded segment A in Figure 3.1, and there are $n - n_j - 1$ slices which cannot be assigned to c_j, corresponding to the shaded segment B in Figure 3.1. This means that the minimum number of different arrangements of the two future observations such that only one is in category c_j is equal to $(n_j - 1)(n - n_j - 1)$. When $0 < n_j < n$, the NPI lower probability is therefore

$$\underline{P}(M_j = 1) = \binom{n+m-1}{m}^{-1} (n_j - 1)(n - n_j - 1).$$

In the case $n_j = 0$, we have $\underline{P}(M_j = 1) = 0$. This is also true when $n_j = n$, since every slice and therefore both future observations may then be assigned to c_j.

We now consider the NPI upper probability for the event $M_j = 1$. There are $n_j + 1$ slices of the wheel which we can assign to category c_j. This includes two optional slices (labelled 1 and 2 in Figure 3.1), which we define as slices that we may or may not assign to c_j. There are a number of different arrangements of the two future observations which can lead to $M_j = 1$. First, we could have one observation in the c_j segment and one in the remainder of the wheel. Secondly, we could have both observations in the c_j segment, with one of these falling in an

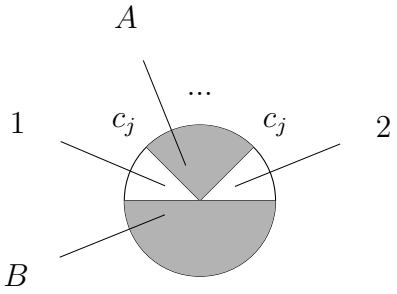

Figure 3.1: Relevant segments and slices of probability wheel for NPI lower and upper probabilities

optional slice. There are $2(n_j - 1)$ such arrangements, since there are two optional slices. Thirdly, we could have both observations in the optional slices. There are three such arrangements, since we could have both observations in the same slice or one in each. This means that the maximum number of different arrangements of the two future observations such that only one is in category c_j is equal to $(n_j + 1)(n - n_j - 1) + 2(n_j - 1) + 3$. When $0 < n_j < n$, the NPI upper probability is therefore

$$\overline{P}(M_j = 1) = \binom{n+m-1}{m}^{-1} [(n_j + 1)(n - n_j - 1) + 2(n_j - 1) + 3].$$

In the cases $n_j = 0$ and $n_j = n$, the derivation of the upper probability is slightly different, because when $n_j = 0$, there is only one slice which we may assign to c_j and when $n_j = n$, there is only one slice which we may assign to a category other than c_j. We therefore have $\overline{P}(M_j = 1) = \binom{n+m-1}{m}^{-1} n$ for $n_j = 0$ and $n_j = n$.

It may also be of interest to consider the probability for the event that both future observations are in category c_j. When considering the lower probability for $M_j = 2$, there are $n_j - 1$ slices of the wheel assigned to observed category c_j. By (3.3), the number of possible arrangements of the two future observations within these slices is equal to $\binom{(n_j-2)+2}{2}$. When $n_j > 1$, the NPI lower probability is therefore

$$\underline{P}(M_j = 2) = \binom{n+m-1}{m}^{-1} \binom{n_j}{2}.$$

In the cases $n_j = 0$ and $n_j = 1$ we have $\underline{P}(M_j = 2) = 0$.

When considering the upper probability for $M_j = 2$, there are $n_j + 1$ slices of the wheel assigned to category c_j. By (3.3), the number of possible arrangements of the two future observations within these slices is equal to $\binom{n_j+2}{2}$. When $n_j < n-1$ the NPI upper probability is therefore

$$\overline{P}(M_j = 2) = \binom{n+m-1}{m}^{-1} \binom{n_j+2}{2}.$$

In the cases $n_j = n-1$ and $n_j = n$ we have $\overline{P}(M_j = 2) = 1$.

Example 3.3.2. Consider the data set described in Example 3.3.1. Suppose that we make inferences about two future observations and we want to select a single category with the largest NPI lower or upper probability for the event that both future observations are in that category. We first consider category B.

To find the NPI lower probabilities, we require a configuration of the probability wheel where B is assigned the minimum possible number of slices, such as the configuration shown in Figure 3.2. We evaluate the NPI lower probability for the

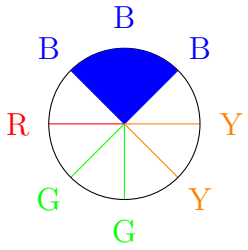

Figure 3.2: Probability wheel for Example 3.3.2 corresponding to lower probability

event that precisely one of the two future observations is in B. This is equal to

$$\underline{P}(M_B = 1) = \binom{n+m-1}{m}^{-1} (n_B - 1)(n - n_B - 1) = \frac{1}{36}(2)(4) = \frac{8}{36}.$$

We also evaluate the NPI lower probability for the event that both future observations are in B. This is equal to

$$\underline{P}(M_B = 2) = \binom{n+m-1}{m}^{-1} \binom{n_B}{2} = \frac{1}{36}\binom{3}{2} = \frac{3}{36}.$$

To find the NPI upper probabilities, we configure the wheel such that two more slices of the wheel may be assigned to B, as shown in Figure 3.3. The NPI upper

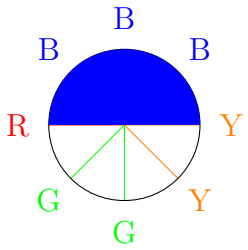

Figure 3.3: Probability wheel for Example 3.3.2 corresponding to upper probability

probability for the event that precisely one of the two future observations is in B is equal to

$$\begin{aligned}\overline{P}(M_B = 1) &= \binom{n+m-1}{m}^{-1} [(n_B+1)(n-n_B-1) + 2(n_B-1) + 3] \\ &= \frac{1}{36}[(4)(4) + 2(2) + 3] = \frac{23}{36}.\end{aligned} \tag{3.4}$$

The NPI upper probability for the event that both future observations are in B is equal to

$$\overline{P}(M_B = 2) = \binom{n+m-1}{m}^{-1} \binom{n_B+2}{2} = \frac{1}{36}\binom{5}{2} = \frac{10}{36}.$$

So $P(M_B = 1) = [\frac{8}{36}, \frac{23}{36}]$ and $P(M_B = 2) = [\frac{3}{36}, \frac{10}{36}]$. We now use the same method to evaluate the equivalent probabilities for the remaining categories. We find that $P(M_G = 1) = P(M_Y = 1) = [\frac{5}{36}, \frac{20}{36}]$ and $P(M_R = 1) = [0, \frac{15}{36}]$ and that $P(M_G = 2) = P(M_Y = 2) = [\frac{1}{36}, \frac{6}{36}]$ and $P(M_R = 2) = [0, \frac{3}{36}]$. This information allows us to select B as the category with the largest NPI lower and upper probability for the event that both future observations are in that category. ◇

m future observations

We now consider the general case where m may take any value. We want to find the probability for the event that some proportion of these m future observations is in category c_j. We may wish to specify a particular number of observations, corresponding to the event of interest $M_j = m_j$ for some number $0 < m_j \leq m$. We may also wish to specify a threshold for M_j, corresponding to the event $M_j \geq m_j$

for some number $m_j \leq m$. The general formulae presented in this subsection hold for $1 < n_j < n - 1$. Other cases are considered separately.

Deriving $\mathbf{P(M_j = m_j)}$

First, we derive the NPI lower probability for the event that precisely m_j of the m future observations belong to category c_j. Figure 3.1 illustrates the two relevant segments, A and B, of the wheel. The shaded segment A in Figure 3.1 represents all slices which must be assigned to c_j. There are $n_j - 1$ such slices. By (3.3), the number of different arrangements of m_j future observations within this segment is equal to

$$\binom{n_j - 2 + m_j}{m_j}. \tag{3.5}$$

The shaded segment B in Figure 3.1 represents all slices which must be assigned to a category other than c_j. There are $n - n_j - 1$ such slices. By (3.3), the number of different arrangements of $m - m_j$ future observations within this segment is equal to

$$\binom{n - n_j - 2 + (m - m_j)}{m - m_j}. \tag{3.6}$$

Multiplying (3.5) and (3.6) gives the minimum number of arrangements in which m_j future observations are in c_j, showing that the NPI lower probability is equal to

$$\underline{P}(M_j = m_j) = \binom{n + m - 1}{m}^{-1} \binom{n_j - 2 + m_j}{m_j} \binom{n - n_j - 2 + (m - m_j)}{m - m_j}. \tag{3.7}$$

This general formula holds for any positive integers m and m_j such that $m_j \leq m$. In the case $n_j \leq 1$ we are not forced to assign any slices of the wheel to c_j, leading to $\underline{P}(M_j = m_j) = 0$.

We now find the corresponding upper probability. We want to maximise the number of arrangements of the m future observations in which m_j future observations are in c_j. As in the case of lower probability, we count all arrangements where m_j observations fall in segment A and $m - m_j$ observations fall in segment B (see Figure 3.1). We showed previously that there are

$$\binom{n_j - 2 + m_j}{m_j} \binom{n - n_j - 2 + (m - m_j)}{m - m_j}$$

such arrangements. However, we now also consider the two optional slices on the wheel (labelled 1 and 2 in Figure 3.1). Any observations which fall in one of the optional slices may be counted either as belonging to c_j or as not belonging to c_j. This means that to find the upper probability we need to include arrangements with one or more observations in the optional slices. Let T denote the total number of future observations in the optional slices, where T ranges from 1 to m. For $T = 1$, there are two possible orderings, as the observation could fall either in slice 1 or in slice 2. By similar reasoning, for $T = 2$, there are three possible orderings. In general, there are $T + 1$ possible orderings for each value of T.

For each value of T there are a number of different arrangements of the m future observations that give T observations in the optional slices. Let X be an integer such that $X \leq m_j$ and $T - X \leq m - m_j$. Then we may have $m_j - X$ observations in segment A, $(m - m_j) - (T - X)$ observations in segment B, and T observations in the optional slices, where X ranges from $T - (m - m_j)$ to m_j. Therefore, the total number of arrangements belonging to the event $M_j = m_j$ with one or more observations in the optional slices is equal to

$$\sum_{T=1}^{m} \sum_{X=(T-(m-m_j))^+}^{\min\{m_j,T\}} (T+1) \binom{n_j - 2 + (m_j - X)}{m_j - X} \binom{n - n_j - 2 + (m - m_j) - (T - X)}{m - m_j - (T - X)}.$$

This enables us to find the maximum number of different arrangements of the m future observations in which m_j observations are in c_j, leading to the NPI upper probability

$$\begin{aligned} \overline{P}(M_j = m_j) = \binom{n + m - 1}{m}^{-1} [& \binom{n_j - 2 + m_j}{m_j} \binom{n - n_j - 2 + (m - m_j)}{m - m_j} \\ & + \sum_{T=1}^{m} \sum_{X=(T-(m-m_j))^+}^{\min\{m_j,T\}} \\ & (T + 1) \binom{n_j - 2 + (m_j - X)}{m_j - X} \binom{n - n_j - 2 + (m - m_j) - (T - X)}{m - m_j - (T - X)}]. \end{aligned} \tag{3.8}$$

Again, this formula holds for any positive integers m and m_j such that $m_j \leq m$. An unobserved category can be assigned at most one slice of the wheel, leading to

$\overline{P}(M_j = m_j) = \binom{n+m-1}{m}^{-1} \sum_{T=m_j}^{m} \binom{n-2+(m-T)}{m-T}$. In the case $n_j = 1$, the formula becomes $\overline{P}(M_j = m_j) = \binom{n+m-1}{m}^{-1} \sum_{T=m_j}^{m} (T+1)\binom{n-3+(m-T)}{m-T}$. When $n_j \geq n-1$, every slice on the wheel may be assigned to that category.

Example 3.3.3. Consider a multinomial data set with possible categories blue (B), red (R), yellow (Y) and green (G). The data are

$$(n_B, n_R, n_Y, n_G) = (2, 1, 1, 1).$$

Suppose that we make inferences about three future observations and we want to find the NPI lower and upper probabilities for the event that precisely two of these are in B. To find the NPI lower probability, we use (3.7) with $m_j = 2$. Using the values $n = 5$, $m = 3$ and $n_j = 2$,

$$\underline{P}(M_j = 2) = \frac{1}{35}\binom{2}{2}\binom{2}{1} = \frac{2}{35}.$$

To find the NPI upper probability, we use (3.8) with $m_j = 2$, which gives

$$\begin{aligned}\overline{P}(M_j = 2) = \frac{1}{35}[&\binom{2}{2}\binom{2}{1} + \sum_{X=0}^{1} 2\binom{2-X}{2-X}\binom{2-(1-X)}{1-(1-X)} + \\ &\sum_{X=1}^{2} 3\binom{2-X}{2-X}\binom{2-(2-X)}{1-(2-X)} + \sum_{X=2}^{2} 4\binom{2-X}{2-X}\binom{2-(3-X)}{1-(3-X)}] \\ = \frac{1}{35}[&2 + 2\binom{2}{2}\binom{1}{0} + 2\binom{1}{1}\binom{2}{1} + 3\binom{1}{1}\binom{1}{0} + 3\binom{0}{0}\binom{2}{1} + 4\binom{0}{0}\binom{1}{0}] \\ = \frac{1}{35}[&2+2+4+3+6+4] = \frac{21}{35}.\end{aligned}$$

So we see that $P(M_j = 2) = [\frac{2}{35}, \frac{21}{35}]$. ◇

Theorem 3.3.2. *For general m, when selecting the category which has the largest NPI lower or upper probability for the event that the category contains all of the future observations, it is optimal to select the category with the greatest number of data observations.*

Proof. We find the category which maximises $\underline{P}(M_j = m)$ and $\overline{P}(M_j = m)$. The general formulae (3.7) and (3.8) can be simplified in the case $m_j = m$, because in

this case $m - m_j = 0$ and also the only possible value of X in the summation is T, leading to $T - X = 0$. This gives

$$\underline{P}(M_j = m) = \binom{n+m-1}{m}^{-1} \binom{n_j - 2 + m}{m}$$

and

$$\overline{P}(M_j = m) = \binom{n+m-1}{m}^{-1} [\binom{n_j - 2 + m}{m} + \sum_{T=1}^{m} \binom{n_j - 2 + (m-T)}{m-T}].$$

The values of n, m and T do not depend on the category selected and since both of these probability formulae are increasing in n_j, it is always optimal to select the category with the largest value of n_j, i.e. the greatest number of data observations.

□

It is also of interest to investigate which values of n_j maximise the NPI lower probability $\underline{P}(M_j = m_j)$. We denote such values by n_j^*. Plotting $\underline{P}(M_j = m_j)$ against values of n_j ranging from 1 to n shows the graph to be unimodal. Intuitively, we expect that this peak will occur near to $n_j = \frac{nm_j}{m}$, because it seems natural that the proportion of the future observations which are in c_j should be similar to the proportion of the data observations which are in c_j. We now formally assess which values of n_j maximise this lower probability by considering the two ratios

$$\frac{\underline{P}(M_j = m_j | n_j)}{\underline{P}(M_j = m_j | n_j + 1)}$$

and

$$\frac{\underline{P}(M_j = m_j | n_j)}{\underline{P}(M_j = m_j | n_j - 1)},$$

where $\underline{P}(M_j = m_j | n_j)$ denotes the lower probability for the event of interest when there are n_j observations in category c_j.

Theorem 3.3.3. *For general m, the value of n_j which maximises $\underline{P}(M_j = m_j)$ is the integer which lies in the interval $[1 + \frac{m_j}{m}(n-3), 2 + \frac{m_j}{m}(n-3)]$.*

Proof. Consider the ratio

$$\frac{\underline{P}(M_j = m_j | n_j)}{\underline{P}(M_j = m_j | n_j + 1)}.$$

We want to find the point at which this ratio becomes greater than 1, as this gives the smallest value of n_j for which $\underline{P}(M_j = m_j)$ is maximised. This ratio is equal to

$$\begin{aligned}\frac{\underline{P}(M_j = m_j|n_j)}{\underline{P}(M_j = m_j|n_j+1)} &= \frac{\binom{n+m-1}{m}^{-1}\binom{n_j-2+m_j}{m_j}\binom{n-n_j-2+(m-m_j)}{m-m_j}}{\binom{n+m-1}{m}^{-1}\binom{n_j-1+m_j}{m_j}\binom{n-n_j-3+(m-m_j)}{m-m_j}} \\ &= \frac{(n_j-1)(n-n_j-2+m-m_j)}{(n_j-1+m_j)(n-n_j-2)}.\end{aligned}$$

We find the smallest value of n_j for which this is greater than 1, as shown below.

$$(n_j-1)(n-n_j-2+m-m_j) > (n_j-1+m_j)(n-n_j-2) \Rightarrow n_j > 1+\frac{m_j}{m}(n-3).$$

Now, consider

$$\frac{\underline{P}(M_j = m_j|n_j)}{\underline{P}(M_j = m_j|n_j-1)}.$$

We want to find the point at which this ratio becomes less than 1, as this gives the largest value of n_j for which $\underline{P}(M_j = m_j)$ is maximised. This ratio is equal to

$$\begin{aligned}\frac{\underline{P}(M_j = m_j|n_j)}{\underline{P}(M_j = m_j|n_j-1)} &= \frac{\binom{n+m-1}{m}^{-1}\binom{n_j-2+m_j}{m_j}\binom{n-n_j-2+(m-m_j)}{m-m_j}}{\binom{n+m-1}{m}^{-1}\binom{n_j-3+m_j}{m_j}\binom{n-n_j-1+(m-m_j)}{m-m_j}} \\ &= \frac{(n_j-2+m_j)(n-n_j-1)}{(n_j-2)(n-n_j-1+m-m_j)}.\end{aligned}$$

We find the smallest value of n_j for which this is less than 1, as shown below.

$$(n_j-2+m_j)(n-n_j-1) < (n_j-2)(n-n_j-1+m-m_j) \Rightarrow n_j > 2+\frac{m_j}{m}(n-3).$$

So we see that $\underline{P}(M_j = m_j|n_j)$ is unimodal as a function of n_j for a given m_j and n_j^* must lie in the interval $[1+\frac{m_j}{m}(n-3), 2+\frac{m_j}{m}(n-3)]$. This interval has width 1, and n_j^* must necessarily be an integer. Excluding the case $m_j = m$ (covered in Theorem 3.3.2), the interval limits are non-integers, so there is only one possible value for n_j^*, this being the integer part of $2+\frac{m_j}{m}(n-3)$. □

To see whether this result corresponds to our initial prediction, we check whether $\frac{nm_j}{m}$ lies within this interval, as shown below.

$$1+\frac{m_j}{m}(n-3) < \frac{nm_j}{m} \Rightarrow m_j > \frac{1}{3}m$$

$$\frac{nm_j}{m} < 1+\frac{m_j}{m}(n-3) \Rightarrow m_j < \frac{2}{3}m$$

We see that if $\frac{1}{3}m < m_j < \frac{2}{3}m$, then $\frac{nm_j}{m}$ is indeed within the interval. We also find that if $m_j < \frac{1}{3}m$, then $\frac{nm_j}{m} + 1$ is within the interval, meaning that $\frac{nm_j}{m}$ is just to the left of the interval. Similarly, if $m_j > \frac{2}{3}m$, then $\frac{nm_j}{m} - 1$ is within the interval, meaning that $\frac{nm_j}{m}$ is just to the right of the interval. So in all cases, the optimal value n_j^* is close to $\frac{nm_j}{m}$, as intuitively expected.

Corollary 3.3.1. *For general m, when selecting a category which maximises $\underline{P}(M_j = m_j)$, the optimal category is selected as follows:*

1. *If there exists c_j such that $n_j \in [1+\frac{m_j}{m}(n-3), 2+\frac{m_j}{m}(n-3)]$, then this category is optimal.*

2. *If there is no c_j such that $n_j \in [1 + \frac{m_j}{m}(n-3), 2 + \frac{m_j}{m}(n-3)]$, then find the value of n_j which is closest to the interval on each side. Compare the values of $\underline{P}(M_j = m_j)$ for the two corresponding categories. The category which gives the largest probability is optimal.*

We also notice that if there are many observations and if both m_j and m are very large and $\frac{m_j}{m}$ tends to some limit l, then the interval $[1 + \frac{m_j}{m}(n-3), 2 + \frac{m_j}{m}(n-3)]$ will shrink to the point value nl. This means that the value of the ratio $\frac{n_j}{n}$ with n_j such that it maximises $\underline{P}(M_j = m_j)$ will tend to the same limit l, as is to be expected.

Note that it is also possible to find the values of n_j that maximise the NPI upper probability $\overline{P}(M_j = m_j)$. This is not shown here, but the method is the same as that used in the proof of Theorem 3.3.3. Example 3.3.4 illustrates how Theorem 3.3.3 and its corollary can be implemented.

Example 3.3.4. Consider a multinomial data set with possible categories blue (B), red (R), yellow (Y) and green (G). The data are

$$(n_B, n_R, n_Y, n_G) = (20, 25, 28, 32).$$

Suppose that we make inferences about the next fifty observations (so $m = 50$) and we want to select the category that maximises $\underline{P}(M_j = 11)$.

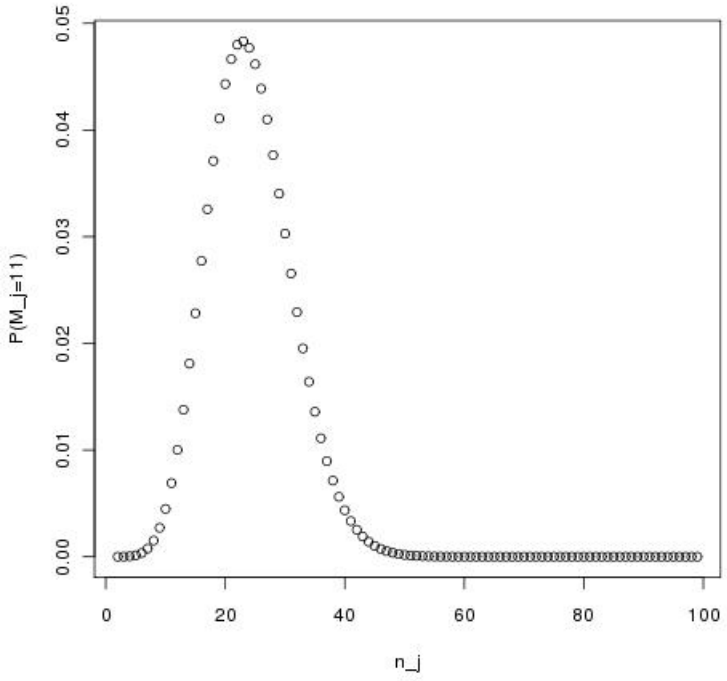

Figure 3.4: Graph of $\underline{P}(M_j = 11)$ against n_j for Example 3.3.4

The plot of $\underline{P}(M_j = 11)$ against all possible values of n_j is shown in Figure 3.4. From this graph, we see that n_j^* will be either 20 or 25, as this is where the peak occurs. By Theorem 3.3.3, the optimal value n_j^* lies in the interval $[1 + \frac{11}{50}(97), 2 + \frac{11}{50}(97)] = [22.34, 23.34]$, so the ideal choice of n_j would be $n_j = 23$. However, there is no c_j in the data set with this value of n_j, so by Corollary 3.3.1 we must look at either side of the interval. To the left of the interval, we have $n_j = 20$ corresponding to B. By (3.7), the relevant lower probability here is $\underline{P}(M_B = 11) = 0.0443$. To the right of the interval, we have $n_j = 25$ corresponding to R. The relevant lower probability here is $\underline{P}(M_R = 11) = 0.0462$. As the second lower probability is largest, we see that $n_j = 25$ is the optimal choice, so we select R as the optimal category. ◇

Deriving $\mathbf{P(M_j \geq m_j)}$

A second event of interest here is the event that at least m_j of the m future

observations are in category c_j. We first consider the NPI lower probability for this event. As before, we count the minimum number of relevant arrangements of the future observations. However, we are now interested in all arrangements which have R future observations that fall in the shaded segment A, where $m_j \leq R \leq m$. We consider each possible value of R separately in order to avoid counting any arrangements more than once. For a given value of R, there are

$$\binom{n_j - 2 + R}{R} \tag{3.9}$$

different arrangements within this segment. We must also consider the remaining $m - R$ observations. Note that, contrary to the lower probability formula above (3.7), arrangements with one or more observations in an optional slice will now be counted. We did not count these when deriving the lower probability $\underline{P}(M_j = m_j)$, because for example an arrangement with m_j observations in segment A and one observation in an optional slice could be allocated to the event $M_j = m_j + 1$ when trying to minimise the probability for the event $M_j = m_j$. However, such arrangements are now relevant because we are simultaneously considering all events $M_j \in \{m_j, m_j + 1, ..., m\}$.

By (3.3), the number of different arrangements of $m - R$ future observations within the shaded segment B plus the two optional slices is equal to

$$\binom{n - n_j + (m - R)}{m - R}. \tag{3.10}$$

Multiplying (3.9) and (3.10) and summing over R from m_j to m gives the NPI lower probability

$$\underline{P}(M_j \geq m_j) = \binom{n + m - 1}{m}^{-1} \sum_{R=m_j}^{m} \binom{n_j - 2 + R}{R} \binom{n - n_j + (m - R)}{m - R}. \tag{3.11}$$

This formula holds for any positive integers m and m_j such that $m_j \leq m$. In the case $n_j \leq 1$ we are not forced to assign any slices of the wheel to c_j, leading to $\underline{P}(M_j \geq m_j) = 0$.

To find the corresponding upper probability, we maximise the number of

arrangements which have at least m_j of the m future observations in category c_j. We still need to count all of the arrangements included for the lower probability, so

$$\sum_{R=m_j}^{m} \binom{n_j - 2 + R}{R} \binom{n - n_j + (m - R)}{m - R} \tag{3.12}$$

arrangements are included in our total. However, we must also include arrangements where there are fewer than m_j observations in segment A but where observations in the optional slices can be counted as belonging to c_j.

Suppose we have Y observations in segment A, where $0 \leq Y \leq m_j - 1$. We need to count any arrangement which has $m_j - Y$ or more observations in an optional slice. Let T denote the total number of future observations in the optional slices. T may range from $m_j - Y$ to $m - Y$ for a given value of Y. As explained above, there are $T + 1$ possible orderings of these observations for each value of T. Therefore, by (3.3), the number of different arrangements is equal to

$$\sum_{Y=0}^{m_j-1} \sum_{T=m_j-Y}^{m-Y} (T+1) \binom{n_j - 2 + Y}{Y} \binom{n - n_j - 2 + (m - Y - T)}{m - Y - T}. \tag{3.13}$$

Summing (3.12) and (3.13) gives the total number of relevant arrangements, leading to the NPI upper probability

$$\overline{P}(M_j \geq m_j) = \binom{n + m - 1}{m}^{-1} [\sum_{R=m_j}^{m} \binom{n_j - 2 + R}{R} \binom{n - n_j + (m - R)}{m - R} + \sum_{Y=0}^{m_j-1} \sum_{T=m_j-Y}^{m-Y} (T+1) \binom{n_j - 2 + Y}{Y} \binom{n - n_j - 2 + (m - Y - T)}{m - Y - T}]. \tag{3.14}$$

Again, this formula is valid for any positive integers m and m_j such that $m_j \leq m$. An unobserved category can be assigned at most one slice of the wheel, leading to $\overline{P}(M_j \geq m_j) = \binom{n + m - 1}{m}^{-1} \sum_{T=m_j}^{m} \binom{n - 2 + (m - T)}{m - T}$. When $n_j = 1$, the formula is $\overline{P}(M_j \geq m_j) = \binom{n + m - 1}{m}^{-1} \sum_{T=m_j}^{m} (T+1) \binom{n - 3 + (m - T)}{m - T}$. When $n_j \geq n - 1$, every slice on the wheel may be assigned to that category.

These formulae can be used in various different ways. For example, suppose that we want to select a category with a NPI lower probability of at least 0.75 for the event that two or more of the future observations will be in that category. We use (3.11) to find all c_j such that $\underline{P}(M_j \geq 2) \geq 0.75$. (Note that if such a category does not exist we may instead select a subset of categories that meets this criterion, as discussed in Section 3.4.) Alternatively, suppose that we want to select a category based on the NPI lower and upper probabilities for the event that the category contains 10% or more of the future observations. We use (3.11) and (3.14) to evaluate $\underline{P}(M_j \geq \frac{m}{10})$ and $\overline{P}(M_j \geq \frac{m}{10})$ for each of the possible categories and then select the category according to these values. This method of selection is illustrated in Example 3.3.5.

Example 3.3.5. Consider the data set described in Example 3.3.3. Suppose that we make inferences about three future observations and we want to select the category with the largest NPI lower and upper probabilities for the event that the category contains at least one of the future observations. To find the NPI lower probability for the event $M_j \geq 1$, we use (3.11) with $m_j = 1$. The first category considered is B. Using the values $n = 5$, $m = 3$ and $n_j = 2$, we find that

$$\underline{P}(M_B \geq 1) = \frac{1}{35}[\binom{5}{2} + \binom{4}{1} + \binom{3}{0}] = \frac{15}{35}.$$

To find the NPI upper probability, we use (3.14) with $m_j = 1$. For B, this gives

$$\begin{aligned}\overline{P}(M_B \geq 1) &= \frac{1}{35}[15 + \sum_{T=1}^{3}(T+1)\binom{0}{0}\binom{4-T}{3-T}] \\ &= \frac{1}{35}[15 + 2\binom{3}{2} + 3\binom{2}{1} + 4\binom{1}{0}] = \frac{31}{35}.\end{aligned}$$

So we see that $P(M_B \geq 1) = [\frac{15}{35}, \frac{31}{35}]$. For all three remaining categories $P(M_j \geq 1) = [0, \frac{25}{35}]$. So the category we select here is B. The key aspect of this inference is the quantification that B contains at least one of the three future observations with lower probability $\frac{15}{35}$. ◇

3.4 NPI subset selection for multinomial data

We now consider the use of NPI to select a subset of categories, rather than a single category, from a multinomial data set. As before, we have K possible categories, where K is a known value, and we have a data set consisting of n observations where n_j denotes the number of times we have observed category c_j for $j = 1, ..., K$. Recall that k represents the total number of categories that have been observed. We select a subset based on inferences about m future observations.

3.4.1 One future observation

We select a subset based on the NPI lower probability for the event that the next observation, Y_{n+1}, belongs to a category within that subset.

Let S denote the selected subset of categories. Let OS denote the index-set for observed categories in S and let US denote the index-set for unobserved categories in S. The sizes of these sets are denoted by r and l respectively. Then, according to the MNPI model described in Chapter 2, it follows from (2.3) and (2.4) that the formula for the lower probability $\underline{P}(Y_{n+1} \in S)$ is

$$\underline{P}(Y_{n+1} \in S) = \sum_{j \in OS} \frac{n_j - 1}{n} + \frac{(2r + l - K)^+}{n} \tag{3.15}$$

and it follows from (2.5) and (2.6) that the formula for the upper probability $\overline{P}(Y_{n+1} \in S)$ is

$$\overline{P}(Y_{n+1} \in S) = \sum_{j \in OS} \frac{n_j - 1}{n} + \frac{\min\{2r + l, k\}}{n}. \tag{3.16}$$

Our objective is to find a subset S such that

$$\underline{P}(Y_{n+1} \in S) \geq p^*$$

for some specified threshold probability p^*. We also want S to be of minimal size. If several such subsets exist, we select one with maximum lower probability. We call such a subset S an optimal subset. This method of subset selection is illustrated in Example 3.4.1.

Example 3.4.1. Consider the data set described in Example 3.3.1. Suppose that we want to find a subset of categories S of minimal size which satisfies the criterion

$$\underline{P}(Y_{n+1} \in S) \geq \frac{3}{8}.$$

As shown in Example 3.3.1, B is the optimal choice when we are selecting a single category and $\underline{P}(M_B = 1) = \frac{2}{8}$. So a subset of size 1 does not satisfy our requirements.

We instead look for a subset of size 2. Consider the subset $S = \{\text{B,G}\}$. Here, $r = 2$ and $l = 0$. (3.15) gives

$$\underline{P}(Y_{n+1} \in \{\text{B,G}\}) = \frac{3-1}{8} + \frac{2-1}{8} + (4-4) = \frac{3}{8}.$$

This satisfies the selection criterion. Applying (3.15) to other possible subsets of size 2 shows that $\frac{3}{8}$ is the largest lower probability that we can achieve with a subset of size 2. So the subset we select is $S = \{\text{B,G}\}$. ◇

Theorem 3.4.1. *When $m = 1$ and we want to select a subset of categories in order to maximise the NPI lower probability that the next observation belongs to a category within that subset, the optimal subset has the property that for all $c_j \in S$, $n_j \geq n_l$ for all $c_l \notin S$.*

Proof. Subset selection is based on the NPI lower probability (3.15). The inclusion of an observed category in S adds $\frac{n_j-1}{n}$ to the first term in (3.15) and may add $\frac{2}{n}$ to the second term. The inclusion of an unobserved category in S adds 0 to the first term and may add $\frac{1}{n}$ to the second term. So we should always include observed categories before unobserved categories. Furthermore, the observed categories which give the largest increase to the NPI lower probability when included in S are those with the largest values of n_j. So it is always optimal that for all $c_j \in S$, $n_j \geq n_l$ for all $c_l \notin S$. □

3.4.2 Multiple future observations

We now consider inferences about multiple future observations. This requires some new notation: let the random quantity M_S represent the number of future

observations that are in S. In terms of the probability wheel, the event $M_S = 1$ means that precisely one future observation falls in any of the slices allocated to S. Let $\overline{S}$ denote the subset of categories not in S, then $\overline{OS}$ denotes the index set for observed categories in $\overline{S}$ and $\overline{US}$ denotes the index set for unobserved categories in $\overline{S}$. We label the respective sizes of these sets $\overline{r}$ and $\overline{l}$.

Based on the MNPI model [17], there are

$$L = \sum_{j \in OS} (n_j - 1) + (2r + l - K)^+ \tag{3.17}$$

slices of the wheel which must be assigned to a category in S and there are

$$\overline{L} = \sum_{j \in \overline{OS}} (n_j - 1) + (2\overline{r} + \overline{l} - K)^+$$

slices of the wheel which cannot be assigned to S.

By considering the difference between the lower and upper probabilities given by the MNPI model [17], we see that there are

$$Q = \min\{2r + l, k\} - (2r + l - K)^+ \tag{3.18}$$

optional slices of the wheel, which we can choose to assign either to S or to $\overline{S}$.

Two future observations

When $m = 2$, it may be of interest to consider the probability for the event that precisely one of the two future observations is in the subset S, i.e. the event $M_S = 1$. We first consider the NPI lower probability for this event. As explained above, there are L slices of the wheel which must be assigned to a category in S and there are $\overline{L}$ slices of the wheel which cannot be assigned to S. Multiplying L and $\overline{L}$ gives the minimum number of different arrangements of the two future observations such that precisely one is in S. By (3.3) there are

$$\binom{n + m - 1}{m} = \binom{n + 1}{2}$$

different arrangements in total, leading to the NPI lower probability

$$\underline{P}(M_S = 1) = \binom{n+1}{2}^{-1} L\overline{L}. \tag{3.19}$$

We now consider the corresponding upper probability and we make use of the optional slices. To find the NPI upper probability, we need to maximise the number of different arrangements of the two future observations which have precisely one observation in S. There are a number of ways in which this situation can arise. First, we may have one observation in a slice of the wheel which must be assigned to S and one in a slice which cannot be assigned to S. As shown above, there are $L\overline{L}$ such arrangements. Secondly, we may have one observation which falls in an optional slice and one observation which falls in any slice other than one of these optional slices. There are $Q(n-Q)$ such arrangements. Finally, we may have both observations in an optional slice. There are

$$\sum_{i=1}^{Q} i = \frac{Q(Q+1)}{2}$$

such arrangements. Summing all of these arrangements leads to the NPI upper probability

$$\overline{P}(M_S = 1) = \binom{n+1}{2}^{-1} \{L\overline{L} + Q(n-Q) + \frac{Q(Q+1)}{2}\}. \tag{3.20}$$

It may also be of interest to consider the probability for the event that both future observations are in the subset S. There are L slices of the wheel that must be assigned to S, so the minimum number of arrangements of the two future observations such that both are in S is equal to

$$\frac{L(L+1)}{2}.$$

This leads to the NPI lower probability

$$\underline{P}(M_S = 2) = \binom{n+1}{2}^{-1} \frac{L(L+1)}{2}. \tag{3.21}$$

In total, there are

$$L + Q = \sum_{j \in OS} (n_j - 1) + \min\{2r + l, k\}$$

slices of the wheel that may be assigned to S. The maximum number of arrangements such that both future observations are in S is equal to

$$\frac{(L+Q)(L+Q+1)}{2}$$

which gives the NPI upper probability

$$\overline{P}(M_S = 2) = \binom{n+1}{2}^{-1} \frac{(L+Q)(L+Q+1)}{2}. \tag{3.22}$$

Example 3.4.2. Consider the data set described in Example 3.3.1. Suppose that we are interested in the subset of categories $S = \{\text{B,G}\}$ and that we make inferences about two future observations. Using (3.19), we evaluate the NPI lower probability for the event that precisely one of the two future observations is in S. We find that $r = \overline{r} = 2$ and $l = \overline{l} = 0$, giving $L = 3$ and $\overline{L} = 1$. This leads to

$$\underline{P}(M_S = 1) = \binom{9}{2}^{-1} L\overline{L} = \frac{1}{36}(3)(1) = \frac{3}{36}.$$

We use (3.21) to find the NPI lower probability for the event that both of the future observations are in S. We have

$$\underline{P}(M_S = 2) = \binom{9}{2}^{-1} \frac{L(L+1)}{2} = \binom{9}{2}^{-1} \frac{3(4)}{2} = \frac{6}{36}.$$

We now consider the NPI upper probabilities. We find that $Q = \min\{4, 4\} - 0 = 4$, and by using (3.20), we find that

$$\overline{P}(M_S = 1) = \binom{9}{2}^{-1} \{(3)(1) + (4)(8-4) + \frac{4(5)}{2}\} = \frac{29}{36}.$$

We also use (3.22) to give

$$\overline{P}(M_S = 2) = \binom{9}{2}^{-1} \frac{(3+4)(3+4+1)}{2} = \frac{28}{36}.$$

◇

A further event of interest is that at least one of the two future observations is in the subset S. In terms of the probability wheel, the event $M_S \geq 1$ means that either one or two future observations fall in any of the slices assigned to S.

To find the NPI lower probability $\underline{P}(M_S \geq 1)$, we must find the minimum number of arrangements of the two future observations such that at least one is in a slice assigned to S. This means that we have to count all arrangements where one observation is in a slice that must be assigned to S, but one observation may be anywhere on the wheel.

There are L slices which must be assigned to S and there are $n - L$ slices which do not have to be assigned to S, leading to $L(n - L)$ arrangements where only one observation is in a slice that must be assigned to S. Also, there are $\frac{L(L+1)}{2}$ arrangements where both observations are in a slice that must be assigned to S. Summing these arrangements leads to the NPI lower probability

$$\underline{P}(M_S \geq 1) = \binom{n+1}{2}^{-1} [L(n-L) + \frac{L(L+1)}{2}]. \tag{3.23}$$

To find the equivalent upper probability, we need to find the corresponding maximum number of arrangements. This means we must count all arrangements where one observation is in a slice that may be assigned to S but one observation can be anywhere on the wheel. There are $L + Q$ slices which may be assigned to S and therefore there are $(L + Q)(n - L - Q)$ arrangements where only one observation is in a slice that must be assigned to S. There are also $\frac{(L+Q)(L+Q+1)}{2}$ arrangements where both observations are in such a slice. Summing these arrangements gives the NPI upper probability

$$\overline{P}(M_S \geq 1) = \binom{n+1}{2}^{-1} [(L+Q)(n-L-Q) + \frac{(L+Q)(L+Q+1)}{2}]. \tag{3.24}$$

Example 3.4.3. Recall Example 3.4.2, where we consider the subset $S = \{\text{B,G}\}$. Suppose that we are now interested in the event $M_S \geq 1$. We previously saw that $L = 3$ and $Q = 4$ for this example. Using (3.23),

$$\underline{P}(M_S \geq 1) = \binom{9}{2}^{-1} [3(5) + \frac{3(4)}{2}] = \frac{21}{36}$$

and using (3.24),

$$\overline{P}(M_S \geq 1) = \binom{9}{2}^{-1} [(7)(1) + \frac{(7)(8)}{2}] = \frac{35}{36}.$$

This implies that there is only one arrangement of the two future observations such that neither observation falls in a slice which may be assigned to S, which means that $\underline{P}(M_S = 0) = \frac{1}{36}$. It is interesting to compare this example to Example 3.4.2, as the superadditivity of the NPI lower probabilities and the subadditivity of the NPI upper probabilities is illustrated. $\diamond$

m future observations

We now consider the general case where m may take any value. We focus on the event that M_S reaches a specified threshold value, i.e. the event $M_S \geq m_S$, because for selection purposes, this is a more natural and useful event to consider than the event that M_S takes one specific value. We assume $0 < L < n$ (see (3.17)), because $L = 0$ leads to lower probability zero. Also, we only have to consider $m_s > 0$ since $\underline{P}(M_S \geq 0) = 1$.

We first consider the NPI lower probability for the event $M_S \geq m_S$. We need to find the minimum number of arrangements of the m future observations such that at least m_S are in the subset S. We do this by counting all arrangements such that R observations fall in slices which must be assigned to S, where $m_S \leq R \leq m$. It is important that we do not count any arrangement multiple times, so we consider each value of R separately and then sum over R to avoid this.

There are L slices which must be assigned to S, so for a certain value of R, there are

$$\binom{L-1+R}{R} \tag{3.25}$$

arrangements of the R observations within the slices which must be assigned to S.

We must also account for the other $m - R$ observations. The remainder of the wheel consists of $n - L$ slices and by (3.3) there are

$$\binom{n-L-1+(m-R)}{m-R} \tag{3.26}$$

different arrangements of the $m - R$ observations within these slices.

Multiplying (3.25) and (3.26) gives the minimum number of arrangements for which $M_S = R$. Summing over all relevant values of R leads to the NPI lower probability

$$\underline{P}(M_S \geq m_s) = \binom{n+m-1}{m}^{-1} \sum_{R=m_S}^{m} \binom{L-1+R}{R} \binom{n-L-1+(m-R)}{m-R}. \tag{3.27}$$

We now consider the NPI upper probability, which means that we need to maximise the number of arrangements which have at least m_S of the m future observations in the subset S. We must still count all of the arrangements described above, i.e. those where at least m_S of the future observations are in slices which must be assigned to S. As explained above, there are

$$\binom{L-1+R}{R} \binom{n-L-1+(m-R)}{m-R}$$

such arrangements.

However, there are other arrangements which must now be included, as we can make use of the Q optional slices (see (3.18)). If we have fewer than m_S observations in slices which must be assigned to S, but we have observations which fall in the Q optional slices, then we can count these observations as belonging to S. It is assumed here that $L + Q < n$. This is because in the situation $L + Q = n$, every slice on the wheel may be assigned to the subset S, leading to the upper probability $\overline{P}(M_S \geq m_S) = 1$.

Suppose that we have Y observations which fall in a slice that must be assigned to the subset S, where $0 \leq Y \leq m_S - 1$. Any arrangement which has $m_S - Y$ or more observations in an optional slice must be counted when calculating the NPI upper probability. Let T denote the total number of future observations in the optional slices. T can take values from $m_S - Y$ to $m - Y$ for a particular value of Y. For each Y, there are

$$\binom{L-1+Y}{Y} \tag{3.28}$$

different arrangements of the Y observations within the slices which must be assigned to S. Also, there are

$$\binom{Q-1+T}{T} \tag{3.29}$$

different arrangements of the T observations within the optional slices. Finally, there are

$$\binom{n-L-Q-1+(m-Y-T)}{m-Y-T} \tag{3.30}$$

different arrangements of the other observations within the remaining slices of the wheel.

Combining (3.28), (3.29) and (3.30) leads to the following NPI upper probability:

$$\overline{P}(M_S \geq m_S) = \binom{n+m-1}{m}^{-1} [\sum_{R=m_S}^{m} \binom{L-1+R}{R}\binom{n-L-1+(m-R)}{m-R} + \sum_{Y=0}^{m_S-1}\sum_{T=m_S-Y}^{m-Y} \binom{L-1+Y}{Y}\binom{Q-1+T}{T}\binom{n-L-Q-1+(m-Y-T)}{m-Y-T}]. \tag{3.31}$$

For $L = 0$, similar arguments directly lead to

$$\overline{P}(M_S \geq m_S) = \binom{n+m-1}{m}^{-1} \times [\sum_{Y=0}^{m_S-1}\sum_{T=m_S-Y}^{m-Y} \binom{Q-1+T}{T}\binom{n-Q-1+(m-Y-T)}{m-Y-T}].$$

Example 3.4.4. Consider the data set described in Example 3.3.3. We make inferences about three future observations and we are interested in the event that at least one of these is in the subset $S = \{\text{B,G}\}$. To find the NPI lower probability for this event, we use (3.27) with $m_S = 1$. We have $n = 5$, $m = 3$,

$$L = \sum_{j \in OS} (n_j - 1) + (2r + l - K)^+ = 1$$

and

$$Q = \min\{2r + l, k\} - (2r + l - K)^+ = 4$$

and so (3.27) gives

$$\underline{P}(M_S \geq 1) = \frac{1}{35}[\binom{1}{1}\binom{5}{2} + \binom{2}{2}\binom{4}{1} + \binom{3}{3}\binom{3}{0}] = \frac{15}{35}.$$

We observe that $L + Q = n$ and this leads to $\overline{P}(M_S \geq 1) = 1$ because we may assign every slice on the wheel to S.

Now suppose that we are interested in the event that at least two of the three future observations are in S. We now apply (3.27) with $m_S = 2$, which gives

$$\underline{P}(M_S \geq 2) = \frac{1}{35}[\binom{2}{2}\binom{4}{1} + \binom{3}{3}\binom{3}{0}] = \frac{5}{35}.$$

As before, every slice on the wheel can be assigned to S, so $\overline{P}(M_S \geq 2) = 1$. So we see that $P(M_S \geq 1) = [\frac{15}{35}, 1]$ and $P(M_S \geq 2) = [\frac{5}{35}, 1]$. ◇

Theorem 3.4.2. *For general m, when selecting a subset of categories in order to maximise the NPI lower probability for the event $M_S \geq m_s$, the optimal subset has the property that for all $c_j \in S$, $n_j \geq n_l$ for all $c_l \notin S$.*

Proof. Our aim is to select the subset which has the highest NPI lower probability

$$\underline{P}(M_S \geq m_s) = \binom{n+m-1}{m}^{-1} \sum_{R=m_s}^{m} \binom{L-1+R}{R}\binom{n-L-1+(m-R)}{m-R}$$

for some given value m_s. L is the only variable in this formula that changes according to which categories are included in S. We therefore wish to determine the behaviour of $\underline{P}(M_S \geq m_s)$ as L increases. To do this, we consider two consecutive values of L. Consider the ratio

$$\frac{\underline{P}(M_S \geq m_s|L)}{\underline{P}(M_S \geq m_s|L+1)}. \tag{3.32}$$

If $\underline{P}(M_S \geq m_s)$ were increasing in L, we would expect this ratio to be always less than 1. Now consider the term within the summation in the formula for this probability. If

$$\frac{\binom{L-1+R}{R}\binom{n-L-1+(m-R)}{m-R}}{\binom{L+R}{R}\binom{n-L+(m-R)}{m-R}} \tag{3.33}$$

is less than 1 for every possible value of R, then (3.32) must always be less than 1. Using the identities of the binomial coefficients, we can rewrite (3.33) as

$$\frac{L(n-L)}{(L+R)(n-L+m-R)}.$$

Then, $L(n-L) < (L+R)(n-L+m-R) \Leftrightarrow 0 < (L+R)(m-R) + R(n-L)$. The term $L+R$ is clearly always positive, $m-R \geq 0$ since $R \leq m$, and $n-L > 0$ since

$L < n$. Therefore $\underline{P}(M_S \geq m_s)$ is increasing in L and our initial aim translates to making L as large as possible.

We now consider how the composition of the subset S affects the value of L. By (3.17), the inclusion of an unobserved category in S adds 0 to the first term in L and may add 1 to the second term in L. The inclusion of an observed category in S adds $n_j - 1$ to the first term in L and may add 2 to the second term in L. So we see that it is optimal to include observed categories in S before unobserved ones. Additionally, we see that the observed categories which increase L by the greatest amount are those with the largest values of n_j. It is therefore optimal that for all $c_j \in S$, $n_j \geq n_l$ for all $c_l \notin S$. □

Example 3.4.5 illustrates how Theorem 3.4.2 can be implemented when selecting subsets.

Example 3.4.5. Consider a multinomial data set with possible categories A to H. The data are shown in Table 3.1. Suppose that we make inferences about three

Category	A	B	C	D	E	F	G	H
Observations	25	20	18	13	10	9	5	0

Table 3.1: Data for Example 3.4.5

future observations and we want to investigate subsets of all possible sizes that maximise the NPI lower probability for the event $M_S \geq m_s$. There are three events of interest here: first, the event that at least one future observation is in some subset S, secondly, the event that at least two future observations are in S and thirdly, the event that all three future observations are in S.

Consider an increasing sequence of subsets $S_1, ..., S_8$, where we begin with a subset of size 1 and add one category at a time. By Theorem 3.4.2, we know that the categories should be added in decreasing order of number of observations. Table 3.2 shows the composition of each of the subsets. Using (3.27) and (3.31)

i	S_i	$P(M_{S_i} \geq 1)$	$P(M_{S_i} \geq 2)$	$P(M_{S_i} \geq 3)$
1	A	$[0.5569, 0.5906]$	$[0.1479, 0.1704]$	$[0.0151, 0.0191]$
2	A,B	$[0.8107, 0.8472]$	$[0.3967, 0.4555]$	$[0.0826, 0.1073]$
3	A-C	$[0.9331, 0.9584]$	$[0.6466, 0.7297]$	$[0.2207, 0.2919]$
4	A-D	$[0.9764, 0.9897]$	$[0.8061, 0.8834]$	$[0.3775, 0.4969]$
5	A-E	$[0.9944, 0.9979]$	$[0.9204, 0.9579]$	$[0.5752, 0.6842]$
6	A-F	$[0.9995, 0.9999]$	$[0.9843, 0.9943]$	$[0.8061, 0.8858]$
7	A-G	$[0.9999, 1.0000]$	$[0.9994, 1.0000]$	$[0.9706, 1.0000]$
8	A-H	$[1.0000, 1.0000]$	$[1.0000, 1.0000]$	$[1.0000, 1.0000]$

Table 3.2: NPI lower and upper probabilities for Example 3.4.5

with $m_S = 1$, we can find the NPI lower and upper probabilities for the event that at least one future observation will be in S_i for $i = 1, ..., 8$. Similarly, we use (3.27) and (3.31) with $m_S = 2$ to find the NPI lower and upper probabilities for the event that at least two future observations will be in S_i for $i = 1, ..., 8$ and we use (3.27) and (3.31) with $m_S = 3$ to find the NPI lower and upper probabilities for the event that all three future observations will be in S_i for $i = 1, ..., 8$. Table 3.2 shows these probabilities.

Suppose that we want to select a subset of minimal size such that the NPI lower probability for the event that one or more of the future observations belongs to a category in that subset is at least 0.5. Looking at Table 3.2 for the event $M_{S_i} \geq 1$, we see that the first S_i which satisfies $\underline{P}(M_{S_i} \geq 1) \geq 0.5$ is $S_1 = \{\text{A}\}$, which we therefore select.

Now suppose that we want to select a subset of minimal size such that the NPI lower probability for the event that two or more of the future observations belong to a category in that subset is at least 0.5. We now need to select a larger subset in order to achieve this minimally required probability. Looking at Table 3.2 for the event $M_{S_i} \geq 2$, we see that the first S_i which satisfies $\underline{P}(M_{S_i} \geq 2) \geq 0.5$ is $S_3 = \{\text{A,B,C}\}$, which we therefore select.

Finally, suppose that we want to select a subset of minimal size such that the NPI lower probability for the event that all future observations belong to a category in that subset is at least 0.5. From Table 3.2 we see that the first S_i which satisfies $\underline{P}(M_{S_i} \geq 3) \geq 0.5$ is $S_5 = \{\text{A,B,C,D,E}\}$, which we therefore select. $\diamond$

3.5 Concluding remarks

In this chapter we presented applications of NPI to selection problems. Methods were presented for selection of a single category and for selection of a subset of categories. These methods incorporated an extension of the MNPI model which enabled inferences about multiple future observations. The results achieved in this chapter could be extended to give NPI lower and upper probabilities for a general event of interest involving multiple future observations and this is an important subject for further research. Such results would make it possible to develop other selection methods using MNPI, e.g. methods analogous to the NPI-based subset selection for real-valued data that was described in Section 3.2. Another possible extension to this work would be the inclusion of observations in the data set which are known to belong to some subset of categories but for which a single category is not specified. This type of missing data is conceptually easy to work with using the MNPI model, as in principle, the lower and upper probabilities for an event of interest can be derived by minimising and maximising over all exact configurations of the wheel that are in line with the available data. The derivation of general formulae could be computationally difficult, however, and the development of selection methods in the case of missing data such as this is a challenging future research topic. A further consideration for future work is the comparison of NPI-based selection methods with other selection methods from the literature. It is clear that the formulation of the inferences given by the MNPI model is very different from that of other methods; for example, indifference zone selection methods give guidance on the sample size, whereas NPI-based selection methods cannot do this. However, a comparison may still lead to useful conclusions.

Chapter 4

Classification

In this chapter we present applications of nonparametric predictive inference (NPI) to classification problems. Throughout this chapter, we assume that K is known (see Subsection 2.2.1) and that there is no ordering of the categories. We adapt the classification tree approach that was outlined in Section 2.3 and we present the use of the MNPI model for building classification trees. As discussed in Section 2.3, we use the maximum entropy distribution when building classification trees using an interval probability model, and in this chapter we present two maximum entropy algorithms for use with the MNPI model. In Section 4.1, we prove that the inferences given by the MNPI model lead to a particular type of interval probability called an F-probability interval. This is useful because it means that the NPI lower and upper probabilities for any general event can be expressed as combinations of the NPI probabilities for single categories. We use the credal set generated by these singleton probabilities as an approximation to the NPI structure and we present an algorithm (A-NPI-M) which finds the maximum entropy distribution within this credal set. In Section 4.2, we consider the true NPI structure $\mathcal{M}_{NPI}$ and we present an algorithm (NPI-M) which finds the maximum entropy distribution within $\mathcal{M}_{NPI}$. In Sections 4.3 and 4.4, we use these algorithms to build classification trees and we compare these methods with other classical and interval probability methods. Finally, in Section 4.5 we consider a variation on the method for building classification trees using the NPI-M algorithm, which involves a bias correction to the Shannon entropy estimator $\widehat{H}$.

4.1 Approximate (A-NPI-M) algorithm

As explained in Section 2.2, the MNPI model can be used to produce an F-probability for the general event $Y_{n+1} \in \bigcup_{j \in J} c_j$ (denoted by E as in (2.2)). For the set of singleton events $Y_{n+1} \in c_j$, $j = 1, ..., K$, the model gives the set of F-probabilities

$$\mathcal{L} = \{[L_j, U_j], j = 1, ..., K\},$$

where $L_j = \underline{P}(Y_{n+1} \in c_j) = \max\{0, \frac{n_j - 1}{n}\}$ and $U_j = \overline{P}(Y_{n+1} \in c_j) = \min\{\frac{n_j + 1}{n}, 1\}$. Recall that k is the number of categories that have been observed. Theorem 4.1.1 shows that the NPI lower and upper probabilities for the general event E can always be determined from the singleton probabilities L_j and U_j. The proof below is based on the configuration of the probability wheel, but Theorem 4.1.1 can also be proven by considering the general formulae for $\underline{P}(E)$ and $\overline{P}(E)$ and this was done by Abellan in parallel to this work for inclusion in a collaborative journal paper [3].

Theorem 4.1.1. *The NPI lower and upper probabilities for E can be derived from the singleton probabilities L_j and U_j, via the following equations:*

$$(a) \quad \underline{P}(E) = \max\{\sum_{j \in J} L_j, 1 - \sum_{j \notin J} U_j\} \tag{4.1}$$

$$(b) \quad \overline{P}(E) = \min\{\sum_{j \in J} U_j, 1 - \sum_{j \notin J} L_j\} \tag{4.2}$$

Proof. (a) We want to show that $\underline{P}(E) = \max\{A, B\}$, where $A = \sum_{j \in J} L_j$ and $B = 1 - \sum_{j \notin J} U_j$. First, we consider the situation where all observed categories in E can be separated on the probability wheel by categories not in E. In this situation, all k separating slices of the wheel can be assigned to E^c and none must be assigned to E. So

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} = \sum_{j \in J} L_j.$$

We have $\sum_{j \notin J} U_j \geq \overline{P}(E^c)$, because $\overline{P}(E^c)$ contains all k separating slices whereas $\sum_{j \notin J} U_j$ includes all k separating slices and may count some of these slices twice. So

therefore, $1 - \sum_{j \notin J} U_j \leq 1 - \overline{P}(E^c)$. By the conjugacy property [17], this means that $\underline{P}(E) \geq 1 - \sum_{j \notin J} U_j$. So we have $\underline{P}(E) = A$ and $\max\{A, B\} = A$. Therefore (4.1) is satisfied.

Secondly, we consider the situation where we cannot separate all observed categories in E on the wheel by categories not in E. In this situation, some of the separating slices must be assigned to E. This means that $\underline{P}(E)$ is necessarily larger than A, since A does not include any separating slices. We see that all observed categories in E^c can be separated on the wheel by categories in E, which means that $\overline{P}(E^c) = \sum_{j \notin J} U_j$. By conjugacy, this implies that

$$\underline{P}(E) = 1 - \sum_{j \notin J} U_j.$$

So we have $\underline{P}(E) = B$ and $\max\{A, B\} = B$. Therefore (4.1) is satisfied.

(b) We want to show that $\overline{P}(E) = \min\{C, D\}$, where $C = \sum_{j \in J} U_j$ and $D = 1 - \sum_{j \notin J} L_j$.

First, we consider the situation where all observed categories in E can be separated on the wheel by observed categories not in E and also $2r + l < k$. In this situation, some of the separating slices have to be assigned to E^c. We assign as many separating slices as possible to E, leading to

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j + 1}{n} + \sum_{j \in UJ} \frac{1}{n} = \sum_{j \in J} U_j.$$

Since we are forced to assign some of the separating slices to E^c, we have $\underline{P}(E^c) > \sum_{j \notin J} L_j$, hence $1 - \sum_{j \notin J} L_j > 1 - \underline{P}(E^c)$. By conjugacy, this implies that $\overline{P}(E) < 1 - \sum_{j \notin J} L_j$. So we have $\overline{P}(E) = C$ and $\min\{C, D\} = C$. Therefore (4.2) is satisfied.

Secondly, we consider the situations where either we cannot separate all observed categories in E on the wheel by observed categories not in E, or we can separate all

observed categories in E but $2r + l \geq k$. In these situations, we are able to assign all k separating slices to E. This means that $\overline{P}(E)$ is necessarily not larger than C, since $\overline{P}(E)$ includes all separating slices whereas C includes all separating slices and may count some of these slices twice. We see that all observed categories in E^c can be separated on the wheel by categories in E, which means that $\underline{P}(E^c) = \sum_{j \notin J} L_j$. By conjugacy, this shows that

$$\overline{P}(E) = 1 - \sum_{j \notin J} L_j.$$

So we have $\overline{P}(E) = D$ and $\min\{C, D\} = D$. Therefore (4.2) is satisfied. □

Note that in general interval probability theory, equations (4.1) and (4.2) do not hold. Interval probabilities which do satisfy these equations have been studied in detail by Weichselberger (see Section 3.3 of [44]) and by De Campos et al. [24]. In Weichselberger's theory, such interval probabilities are termed F-probability intervals, whilst De Campos et al. define these as reachable probability intervals.

The set of F-probability intervals $\mathcal{L}$ is associated with a credal set $\mathcal{P}(\mathcal{L})$ of probability distribution functions, p, which is defined as follows:

$$\mathcal{P}(\mathcal{L}) = \{p |\, j \in \{1, ..., K\},\, p(c_j) \in [L_j, U_j],\, \sum_{j=1}^{K} p(c_j) = 1\}. \tag{4.3}$$

When working with F-probability intervals, the maximum entropy algorithm presented by Abellan and Moral [2] can be used to find the maximum entropy distribution within the associated credal set of probability distributions. The A-NPI-M algorithm presented below is based on this algorithm, but applies specifically to the set $\mathcal{P}(\mathcal{L})$ linked to the MNPI model.

Consider the set $J(t) = \{j | n_j = t\}$ and let $K(t) = |J(t)|$, such that $\sum_{i=0}^{n} K(i) = K$ and $n = \sum_{i=0}^{n} i \cdot K(i)$. Let $K' = K - (K(0) + K(1))$. The algorithm shown below attains the array $\widehat{p} = (\widehat{p}(c_1), \ldots, \widehat{p}(c_K)) \equiv (\widehat{p}_1, \ldots, \widehat{p}_K)$, which is the maximum entropy distribution within the set $\mathcal{P}(\mathcal{L})$. Initially, each category is assigned its

lower probability L_j, then these probabilities are augmented successively beginning with the categories observed least often. The algorithm has been programmed in Weka for use in practical applications, but is written here in pseudo-code. The arrow symbol $\leftarrow$ indicates a new value being assigned to the quantity on the left hand side. Other notation is self-explanatory.

A-NPI-M

If $K' < K(0)$

For $j = 1$ to K

If $(n_j = 0 \ or \ n_j = 1)$ $\widehat{p}_j \leftarrow \frac{K'+K(1)}{n(K(0)+K(1))}$;

Else $\widehat{p}_j \leftarrow \frac{n_j-1}{n}$;

Else

$mass \leftarrow K' - K(0)$;

For $j = 1$ to K

If $(n_j = 0 \ or \ n_j = 1)$ $\widehat{p}_j \leftarrow \frac{1}{n}$;

Else $\widehat{p}_j \leftarrow \frac{n_j-1}{n}$;

$i \leftarrow 1$;

While $(mass > 0)$ do

If $(K(i) + K(i+1) < mass)$

For $j = 1$ to K

If $(n_j = i \ or \ n_j = i+1)$ $\widehat{p}_j \leftarrow \widehat{p}_j + \frac{1}{n}$;

$mass \leftarrow mass - 1$;

Else

For $j = 1$ to K

If $(n_j = i \ or \ n_j = i+1)$ $\widehat{p}_j \leftarrow \widehat{p}_j + \frac{mass}{n(K(i)+K(i+1))}$;

$mass \leftarrow 0$;

$i \leftarrow i + 1$;

It was proven by Abellan and Moral [2] that their algorithm for general credal sets always attains the maximum entropy distribution. This same proof can also be

used to show that the A-NPI-M algorithm attains the maximum entropy distribution within $\mathcal{P}(\mathcal{L})$ (see Theorem 4.1.2 below). Lemmas 4.1.1 and 4.1.2 are needed for this proof. Lemma 4.1.1 was proven by Abellan and Moral [2] and Lemma 4.1.2 was proven by Wasserman and Kadane [42].

Lemma 4.1.1. *Suppose that $\{p_j\}_{j=1}^K$ is an array of K non-negative real numbers and let $q_j = p_j + \epsilon_j$ where $\epsilon_j \geq 0$ for all j. Let $\{p_j^*\}_{j=1}^K$ and $\{q_j^*\}_{j=1}^K$ denote the rearrangements of these arrays such that the values are in decreasing order. Then, $p_j^* \leq q_j^*$ for all $j \in \{1, ..., K\}$.*

Lemma 4.1.2. *Let p and q be two probability distributions on the finite set of categories $c_1, ..., c_K$. Let p^* and q^* be the corresponding arrays reordered in decreasing order. If $\sum_{j=1}^{i} p^*(c_j) \leq \sum_{j=1}^{i} q^*(c_j)$ for $i = 1, ..., K$, then $H(p) \geq H(q)$, where H denotes the Shannon entropy function.*

Theorem 4.1.2. *The probability distribution $\widehat{p}$ that is attained by the A-NPI-M algorithm gives the maximum entropy value of any distribution within the set $\mathcal{P}(\mathcal{L})$.*

Proof. Rearranging the array $\{\widehat{p_j}\}_{j=1}^K$ in decreasing order to give $\widehat{p_K^*} \leq ... \leq \widehat{p_1^*}$, for some s and t we have

$$\widehat{p_j^*} = L_j, j \in \{1, ..., s\},$$

$$L_j < \widehat{p_j^*} = \alpha < U_j, j \in \{s+1, ..., t\}$$

and

$$\widehat{p_j^*} = U_j, j \in \{t+1, ..., K\}.$$

It is sufficient to prove that $\widehat{p}$ is the maximum entropy distribution in $B(\widehat{p}, \epsilon) \cap \mathcal{P}(\mathcal{L})$ for some $\epsilon > 0$, where $B(\widehat{p}, \epsilon)$ is the set of all distributions $\{q_j\}_{j=1}^K$ such that $d(\widehat{p}, q) \leq \epsilon$ for a distance function d on $\mathbb{R}^n$. This is sufficient because H is a convex function [2]. We have

$$q = (\widehat{p}_1 + \epsilon_1, ..., \widehat{p}_s + \epsilon_s, \widehat{p}_{s+1} \pm \epsilon_{s+1}, ..., \widehat{p}_t \pm \epsilon_t, \widehat{p}_{t+1} - \epsilon_{t+1}, ..., \widehat{p}_K - \epsilon_K),$$

where $0 \leq \epsilon_j \leq \epsilon$ for all j. Rearranging in decreasing order, this array becomes

$$q^* = (q_1^*, ..., q_{s_1}^*, q_{s_1+1}^*, ..., q_{t_1}^*, q_{t_1+1}^*, ..., q_K^*),$$

where $s_1 \geq s$, $K - t_1 \geq K - t$ and $q^*_{s_1+1} = ... = q^*_{t_1} = \alpha$. Let ϵ^*_j denote the values corresponding to q^*_j for $j = 1, ..., K$. Since all the terms in the array must sum to 1, we have $\sum_{j=1}^{s_1} \epsilon^*_j = \sum_{j=t_1+1}^{K} \epsilon^*_j$.

Using Lemma 4.1.1, we see that:

- $q^*_j \geq \widehat{p^*_j}$ for $j \in \{1, ..., s\}$, so $\sum_{j=1}^{h} q^*_j \geq \sum_{j=1}^{h} \widehat{p^*_j}$ for all $h \leq s$.
- $q^*_j \geq \alpha = \widehat{p^*_j}$ for $j \in \{s+1, ..., s_1\}$, so $\sum_{j=1}^{h} q^*_j \geq \sum_{j=1}^{h} \widehat{p^*_j}$ for all $h \leq s_1$.
- $q^*_j = \alpha = \widehat{p^*_j}$ for $j \in \{s_1+1, ..., t_1\}$, so $\sum_{j=1}^{h} q^*_j \geq \sum_{j=1}^{h} \widehat{p^*_j}$ for all $h \leq t_1$.

Since $\sum_{j=1}^{t_1} q^*_j = \sum_{j=1}^{t_1} \widehat{p^*_j} + \sum_{j=1}^{s_1} \epsilon^*_j$, and since $\sum_{j=1}^{s_1} \epsilon^*_j = \sum_{j=t_1+1}^{K} \epsilon^*_j$, we must have $\sum_{j=1}^{t_1+h} q^*_j \geq \sum_{j=1}^{t_1+h} \widehat{p^*_j}$ for all $h \in \{1, ..., K - t_1\}$. Therefore, by Lemma 4.1.2, $H(\widehat{p^*}) \geq H(q^*)$, so $H(\widehat{p}) \geq H(q)$. □

The use of the A-NPI-M algorithm for classification is considered in Section 4.3.

4.2 Exact (NPI-M) algorithm

It is important to note that the set of probability distributions generated by the MNPI model, i.e. the NPI structure $\mathcal{M}_{NPI}$, is not a credal set. As shown in Section 4.1, we can determine bounds on the probability for a general event via the NPI probabilities for the single categories, but not all distributions in the associated credal set $\mathcal{P}(\mathcal{L})$ are compatible with the theoretical MNPI model. Example 4.2.1 illustrates this by giving a distribution which is in the set $\mathcal{P}(\mathcal{L})$ but not in $\mathcal{M}_{NPI}$.

Example 4.2.1. Consider a multinomial data set with observed categories R, B and O and unobserved categories U1, U2, U3 and U4. The data are

$$(n_R, n_B, n_O) = (3, 3, 2).$$

According to the MNPI model, the set of probability intervals $\mathcal{L}$ for the set of categories {R,B,O,U1,U2,U3,U4} is given by

$$\{[\frac{2}{8},\frac{4}{8}],[\frac{2}{8},\frac{4}{8}],[\frac{1}{8},\frac{3}{8}],[0,\frac{1}{8}],[0,\frac{1}{8}],[0,\frac{1}{8}],[0,\frac{1}{8}]\}.$$

The maximum entropy probability distribution within the credal set

$$\mathcal{P}(\mathcal{L}) = \{p|\, j \in \{1,...,K\},\, p(c_j) \in [L_j, U_j],\, \sum_{j=1}^{K} p(c_j) = 1\}$$

is given by

$$\widehat{p} = \{\frac{2}{8},\frac{2}{8},\frac{1}{8},\frac{3}{32},\frac{3}{32},\frac{3}{32},\frac{3}{32}\}.$$

However, it is not possible to find a configuration of the probability wheel that corresponds to this distribution and that is in line with the MNPI model. In order to

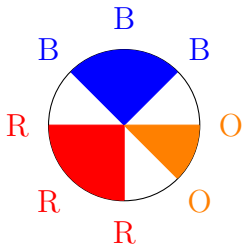

Figure 4.1: Probability wheel for Example 4.2.1

achieve this distribution, the three separating slices of the wheel shown in Figure 4.1 would have to be shared evenly between the four unobserved categories. This would mean that at least one of the unobserved categories would have to be represented by multiple segments of the wheel, which is not allowed by the MNPI model. ◇

The NPI-M algorithm presented below gives the maximum entropy distribution within the true NPI structure $\mathcal{M}_{NPI}$ rather than approximating the NPI structure by a credal set. It is clear that the discrepancy between $\mathcal{M}_{NPI}$ and the credal set $\mathcal{P}(\mathcal{L})$ is due to limitations caused by the configuration of the probability wheel, so the NPI-M algorithm is constructed by considering how best to assign each slice of the wheel. The distribution returned by the algorithm, $p_{maxE} = (p_{maxE}(c_1),\ldots,p_{maxE}(c_K))$, gives the largest entropy value possible whilst still corresponding to a valid configuration of the probability wheel. The cases $K(0) > K'$

and $K(0) \leq K'$ are considered separately, because when $K(0) > K'$ all separating slices of the wheel can be assigned to unobserved categories or to categories observed only once, which makes the problem simpler than in the case $K(0) \leq K'$.

4.2.1 NPI-M algorithm: $K(0) > K'$

As before, let $K(0)$ represent the number of unobserved categories, $K(1)$ the number of categories observed once and K' the number observed twice or more. We first consider situations where $K(0) > K' > 0$. In the trivial case of $K' = 0$, every category is either unobserved or observed only once, so all lower probabilities L_j are zero and upper probabilities U_j are $\frac{1}{n}$ for unobserved categories and $\frac{2}{n}$ for categories observed once. In this case, the uniform probability distribution assigning probability $\frac{1}{K}$ to every category maximises entropy and is in $\mathcal{M}_{NPI}$.

The construction of the maximum entropy distribution presented in this subsection is based on the principle of Lemma 4.1.2. As in the A-NPI-M algorithm described in Section 4.1, categories observed more than once are assigned their lower probability L_j and the remaining probability mass is then shared between the unobserved categories and the categories observed only once. However, there are restrictions on the way in which this can be shared. We need to share the $K' + K(1)$ separating slices between the $K(0) + K(1)$ categories that have a lower probability of zero, in such a way that the resulting distribution gives the largest entropy value possible but does not violate the rules of the MNPI model.

The configuration of the wheel which gives the most flexibility with respect to sharing out the separating slices is the arrangement where all categories that have been observed only once are placed next to each other on the wheel. This results in one segment made up of $K(1)+1$ separating slices, plus $K'-1$ individual separating slices.

As K, K' and $K(i)$, for each i, are integers, we can denote

$$\beta = (K(0) + K(1))/(K' + K(1)) \tag{4.4}$$

and

$$h = (K(0) + K(1))\%(K' + K(1)), \tag{4.5}$$

where "/" represents the integer division operator and "%" represents the remainder operator (also known as the modulo operator).

To find a probability distribution that gives the largest entropy value possible, the categories observed more than once should only be assigned their lower probabilities. By Lemma 4.1.2, an increase in the probability for a category c_j with $n_j \geq 2$ causes a decrease in entropy. The remaining probability mass $\frac{K'+K(1)}{n}$ must be distributed as equally as possible amongst the $K(0) + K(1)$ categories.

As a first step, we assign β (see (4.4)) categories to each of the $K' + K(1)$ separating slices. This is clearly optimal with regard to maximising entropy, since we assign the same number of categories to each available slice. As a second step, we then distribute the remaining h (see (4.5)) categories. We consider the situations $h < K(1) + 1$ and $h \geq K(1) + 1$ separately.

If $h < K(1) + 1$, we assign the h categories to the segment of $K(1) + 1$ separating slices. Due to the constraints of the wheel, this gives more flexibility with respect to sharing out the remaining probability mass than if we were to assign these h categories to single separating slices. So, in this case, we have $K' - 1$ single separating slices that are assigned β categories each and a segment of $K(1) + 1$ separating slices that has $\beta(K(1) + 1) + h$ categories in total. By Lemma 4.1.2, we could only increase entropy by sharing the final h categories between more than $K(1) + 1$ slices, but this is not possible due to the model constraints.

If $h \geq K(1) + 1$, the best way to distribute the remaining mass is to assign each of the h categories to a different separating slice. So, in this case, we have h separating slices that are assigned $\beta + 1$ categories each and $K' + K(1) - h$ separating slices that are assigned β categories each. By Lemma 4.1.2, we could only increase entropy by splitting one or more categories over separate slices, but

again this is not allowed by the model constraints.

For simplicity, suppose that we reorder the categories and label them $c_1, ..., c_K$ in order of increasing n_j. The above method of filling the slices of the wheel then leads to the following maximum entropy probability distribution:

If $\mathbf{h < K(1) + 1}$:

$p_{maxE}(c_j) = \frac{1}{n\beta}$ for $j = 1, ..., \beta(K' - 1)$

$p_{maxE}(c_j) = \frac{K(1)+1}{n[\beta(K(1)+1)+h]}$ for $j = \beta(K' - 1) + 1, ..., K(0) + K(1)$

$p_{maxE}(c_j) = \frac{n_j - 1}{n}$ for $j = K(0) + K(1) + 1, ..., K$

If $\mathbf{h \geq K(1) + 1}$:

$p_{maxE}(c_j) = \frac{1}{n(\beta+1)}$ for $j = 1, ..., h(\beta + 1)$

$p_{maxE}(c_j) = \frac{1}{n\beta}$ for $j = h(\beta + 1) + 1, ..., K(0) + K(1)$

$p_{maxE}(c_j) = \frac{n_j - 1}{n}$ for $j = K(0) + K(1) + 1, ..., K$

In some cases, categories with the same n_j value are assigned different probabilities. With regard to maximising entropy for the purpose of building classification trees, it is irrelevant which of these categories is assigned a larger probability and which is assigned a smaller probability. This is because when calculating the entropy in order to select a split variable (see Subsection 2.3.2) we sum over all categories. The following examples illustrate this method of assigning probabilities. Examples 4.2.2 and 4.2.3 illustrate the situation $h < K(1) + 1$ and Example 4.2.4 illustrates the situation $h \geq K(1) + 1$.

Example 4.2.2. Consider a multinomial data set with observed categories R, B, G, P and O and unobserved categories U1, U2, U3 and U4. The data are

$$(n_R, n_B, n_G, n_P, n_O) = (2, 2, 2, 1, 1).$$

In order of increasing n_j, the set of possible categories is

$$\{\text{U1,U2,U3,U4,P,O,R,B,G}\}.$$

The NPI-M algorithm assigns the lower probability $\frac{n_j-1}{n}$ to categories R, B and G, so $p_{maxE}(\mathrm{R}) = p_{maxE}(\mathrm{B}) = p_{maxE}(\mathrm{G}) = \frac{1}{8}$. We have $K(0) + K(1) = 6$ and $K' + K(1) = 5$, therefore

$$\beta = (K(0) + K(1))/(K' + K(1)) = 1$$

and

$$h = (K(0) + K(1))\%(K' + K(1)) = 1.$$

Here $K(1) = 2$ and so $h < K(1) + 1$. As $\beta(K' - 1) = 2$, two of the unobserved

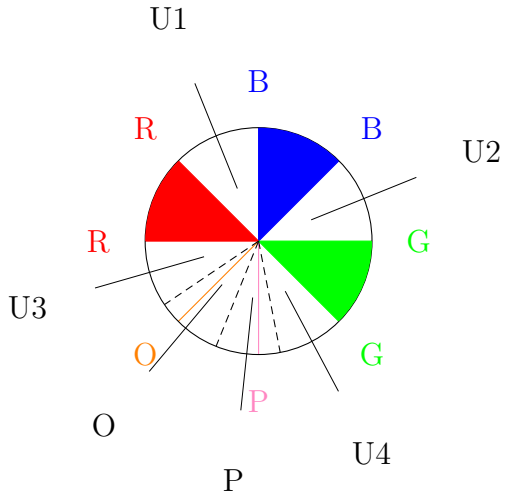

Figure 4.2: Probability wheel for Example 4.2.2

categories, say U1 and U2, are assigned the probability

$$p_{maxE}(c_j) = \frac{1}{n\beta} = \frac{1}{8},$$

while the other two unobserved categories (U3 and U4) and the two categories observed only once (P and O) are assigned the probability

$$p_{maxE}(c_j) = \frac{K(1) + 1}{n[\beta(K(1) + 1) + h]} = \frac{3}{32}.$$

Note that the order within the unobserved categories does not matter. This is because it does not matter which categories are assigned the larger probability, since we sum over all categories when calculating the entropy. A possible configuration of the probability wheel is shown in Figure 4.2. The resulting maximum entropy distribution for the set of categories

$$\{\mathrm{U1,U2,U3,U4,P,O,R,B,G}\}$$

is

$$p_{maxE} = \{\frac{1}{8}, \frac{1}{8}, \frac{3}{32}, \frac{3}{32}, \frac{3}{32}, \frac{3}{32}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\}.$$

◇

Example 4.2.3. Consider a multinomial data set with observed categories R, B, G, P and O and unobserved categories U1 to U9. The data are

$$(n_R, n_B, n_G, n_P, n_O) = (2, 2, 2, 1, 1).$$

In order of increasing n_j, the set of possible categories is

$$\{\text{U1,U2,U3,U4,U5,U6,U7,U8,U9,P,O,R,B,G}\}.$$

The NPI-M algorithm assigns the lower probability $\frac{n_j - 1}{n}$ to categories R, B and G, so $p_{maxE}(\text{R}) = p_{maxE}(\text{B}) = p_{maxE}(\text{G}) = \frac{1}{8}$. We have $K(0) + K(1) = 11$ and $K' + K(1) = 5$, therefore

$$\beta = (K(0) + K(1))/(K' + K(1)) = 2$$

and

$$h = (K(0) + K(1))\%(K' + K(1)) = 1.$$

Here $K(1) = 2$ and so $h < K(1) + 1$. As $\beta(K' - 1) = 4$, four of the unobserved categories, say U1 to U4, are assigned the probability

$$p_{maxE}(c_j) = \frac{1}{n\beta} = \frac{1}{16},$$

while the other five unobserved categories (U5 to U9) and the two categories observed only once are assigned the probability

$$p_{maxE}(c_j) = \frac{K(1) + 1}{n[\beta(K(1) + 1) + h]} = \frac{3}{56}.$$

Again, the order within unobserved categories is not relevant as it does not matter which of these categories is assigned the larger probability. A possible configuration of the probability wheel is shown in Figure 4.3. The resulting maximum entropy distribution for the set of categories

$$\{\text{U1,U2,U3,U4,U5,U6,U7,U8,U9,P,O,R,B,G}\}$$

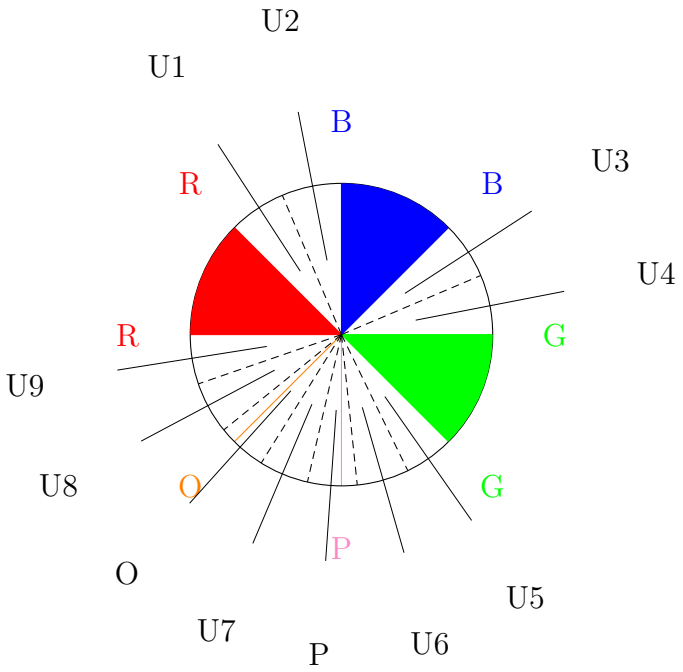

Figure 4.3: Probability wheel for Example 4.2.3

is

$$p_{maxE} = \{\frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{3}{56}, \frac{3}{56}, \frac{3}{56}, \frac{3}{56}, \frac{3}{56}, \frac{3}{56}, \frac{3}{56}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\}.$$

◇

Example 4.2.4. Consider the data set described in Example 4.2.1. In order of increasing n_j, the set of categories is

$$\{\text{U1,U2,U3,U4,O,R,B}\}.$$

The NPI-M algorithm assigns the lower probability $\frac{n_j-1}{n}$ to categories R, B and O, so $p_{maxE}(\text{R}) = p_{maxE}(\text{B}) = \frac{1}{4}$ and $p_{maxE}(\text{O}) = \frac{1}{8}$. We have $K(0) + K(1) = 4$ and $K' + K(1) = 3$, therefore

$$\beta = (K(0) + K(1))/(K' + K(1)) = 1$$

and

$$h = (K(0) + K(1))\%(K' + K(1)) = 1.$$

Here $K(1) = 0$ and so $h \geq K(1) + 1$. As $h(\beta + 1) = 2$, two of the unobserved

categories, say U1 and U2, are assigned the probability

$$p_{maxE}(c_j) = \frac{1}{n(\beta+1)} = \frac{1}{16},$$

while the other two unobserved categories (U3 and U4) are assigned the probability

$$p_{maxE}(c_j) = \frac{1}{n\beta} = \frac{1}{8}.$$

Again, the order within the unobserved categories does not matter. A possible configuration of the probability wheel is shown in Figure 4.4. The resulting

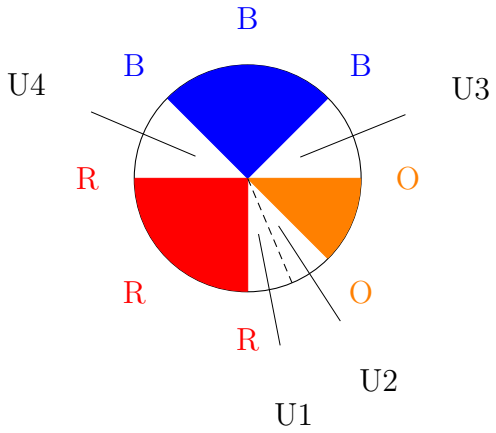

Figure 4.4: Probability wheel for Example 4.2.4

maximum entropy distribution for the set of categories

$$\{\text{U1,U2,U3,U4,O,R,B}\}$$

is

$$p_{maxE} = \{\frac{1}{16}, \frac{1}{16}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{4}\}.$$

◇

The maximum entropy distribution given in Subsection 4.2.1 should only be used when $K(0) > K'$. When $K(0) \leq K'$, the NPI-M algorithm given in Subsection 4.2.2 should be used.

4.2.2 NPI-M algorithm: $K(0) \leq K'$

We now consider situations where $K(0) \leq K'$. As in the A-NPI-M algorithm described in Subsection 4.1, unobserved categories and categories observed only

once are initially assigned probability $\frac{1}{n}$ and categories observed more than once are initially assigned their lower probability L_j. The remaining probability mass is then shared out in such a way that the distribution returned by the algorithm gives the largest entropy value possible with the wheel constraints satisfied and $p_{maxE}(c_j) \leq U_j$ for all categories.

Once the initial probability assignments are made, there are $K' - K(0)$ separating slices remaining. These separating slices are then shared between the categories with the least probability mass, provided that the resulting probabilities are no larger than their upper limits U_j. At each stage, if the number of categories with the least probability mass is smaller than the number of separating slices remaining, each category is assigned one whole slice. Otherwise, the remaining separating slices are evenly divided between as many of the categories as possible. In some cases, where the number of categories with the least probability mass is much larger than the number of separating slices remaining, we cannot share the slices between all relevant categories due to restrictions imposed by the configuration of the probability wheel.

This method of filling the slices of the wheel leads to the following algorithm:

NPI-M $(K(0) \leq K')$

$mass \leftarrow K' - K(0)$;

For $j = 1$ to K

If $(n_j = 0$ *or* $n_j = 1)$ $p_{maxE}(c_j) \leftarrow \frac{1}{n}$;

Else $p_{maxE}(c_j) \leftarrow \frac{n_j - 1}{n}$;

$i \leftarrow 1$;

While $(mass > 0)$ do

If $(K(i) + K(i+1) < mass)$

For $j = 1$ to K

If $(n_j = i$ *or* $n_j = i + 1)$ $p_{maxE}(c_j) \leftarrow p_{maxE}(c_j) + \frac{1}{n}$;

$mass \leftarrow mass - (K(i) + K(i+1))$;

Else

$W \leftarrow \min\{mass + 1 + K(i), K(i) + K(i+1)\}$;

$Acc \leftarrow W$;

For $j = 1$ to K

If $((n_j = i \text{ or } n_j = i+1) \text{ and } (Acc > 0))$

$p_{maxE}(c_j) \leftarrow p_{maxE}(c_j) + \frac{mass}{nW}$;

$Acc \leftarrow Acc - 1$;

$mass \leftarrow 0$;

$i \leftarrow i + 1$;

Theorem 4.2.1. *The probability distribution p_{maxE} that is attained by the NPI-M algorithm for $K(0) \leq K'$ gives the maximum entropy value of any distribution within the NPI structure $\mathcal{M}_{NPI}$.*

Proof. Suppose that we rearrange the array $\{p_{maxE}(c_j)\}_{j=1}^{K}$ in decreasing order. This gives the array

$$p^*_{maxE} = (L_{j_1}, ..., L_{j_R}, Z + \frac{mass_f}{nW}, ..., Z + \frac{mass_f}{nW}, Z, ..., Z, U_{j'_1}, ..., U_{j'_T}),$$

where L_{j_i} represents the NPI lower probability for category c_{j_i}, $U_{j'_i}$ represents the NPI upper probability for category $c_{j'_i}$ and Z represents the probability mass already assigned to categories observed i or $i+1$ times when we reach the final step of the algorithm. At this final step, remaining probability mass ($\frac{mass_f}{n}$) is shared equally between W categories. This is the maximum number of categories for which we can increase the probability from Z, due to the model constraints. Let I denote the number of categories that are assigned a probability such that $L_j < p_{maxE}(c_j) < U_j$. Then $I - W$ categories have probability Z, whilst W categories have probability $Z + \frac{mass_f}{nW}$. Hence, $K = R + I + T$.

Based on Lemma 4.1.2, we want to show that for every distribution q in the NPI structure, $\sum_{j=1}^{i} p^*_{maxE}(c_j) \leq \sum_{j=1}^{i} q^*(c_j)$ for all $i \in \{1, ..., K\}$, where q^* is a rearrangement of the array q in decreasing order.

- It is clear that decreasing any of the elements equal to $L_{j_i}, i = 1, .., R$, in the array p^*_{maxE}, or increasing any of the elements equal to $U_{j'_i}, i = 1, .., T$, does not lead to a valid distribution. Also, if we increase one or more of the L_{j_i} values or decrease one or more of the $U_{j'_i}$ values, we obtain a probability distribution with a smaller entropy than the distribution p_{maxE}.

- If any one of the W elements equal to $Z + \frac{mass_f}{nW}$ is decreased, this will necessitate an increase in one or more of the other W elements or in one of the elements equal to $L_{j_i}, i = 1, .., R$, resulting in a smaller entropy. Note that it is not possible to counteract such a decrease by an increase in one of the $I - W$ elements equal to Z, because the model constraints mean that W is the maximum number of categories for which we can increase the probability from Z.

- If any one of the $I - W$ elements equal to Z is decreased, this will necessitate an increase in one or more of the W elements equal to $Z + \frac{mass_f}{nW}$ or in one or more of the L_{j_i} values, resulting in a smaller entropy.

Therefore, for every distribution q in the NPI structure, $\sum_{j=1}^{i} p^*_{maxE}(c_j) \leq \sum_{j=1}^{i} q^*(c_j)$ for all $i \in \{1, ..., K\}$, so the distribution p_{maxE} must give the maximum entropy value of any distribution within $\mathcal{M}_{NPI}$. □

Example 4.2.5. Consider a multinomial data set with observed categories R, B, G, P, Y, M and O and unobserved categories U1 and U2. The data are

$$(n_R, n_B, n_G, n_P, n_Y, n_M, n_O) = (6, 6, 3, 2, 1, 1, 1).$$

The NPI-M algorithm initially assigns probability masses $\frac{n_j - 1}{n}$ to categories R, B, G and P, which are equal to the corresponding lower probabilities. So $p(\text{R}) = p(\text{B}) = \frac{5}{20}$, $p(\text{G}) = \frac{2}{20}$ and $p(\text{P}) = \frac{1}{20}$. The algorithm initially assigns probability $\frac{1}{20}$ to the remaining categories.

We now reach the recursive part of the algorithm. For $i = 1$, $K(1) + K(2) = 4$ and

$mass = 2$, therefore $K(i) + K(i+1) > mass$ and

$$W = \min\{mass + 1 + K(1), K(1) + K(2)\} = K(1) + K(2) = 4.$$

This means that all four categories which have been observed either once or twice are assigned the probability

$$p_{maxE}(c_j) = \frac{1}{20} + \frac{mass}{20(\min\{mass + 1 + K(1), K(1) + K(2)\})} = \frac{1}{20} + \frac{2}{4 \times 20} = \frac{3}{40}.$$

A possible configuration of the probability wheel is shown in Figure 4.5. The

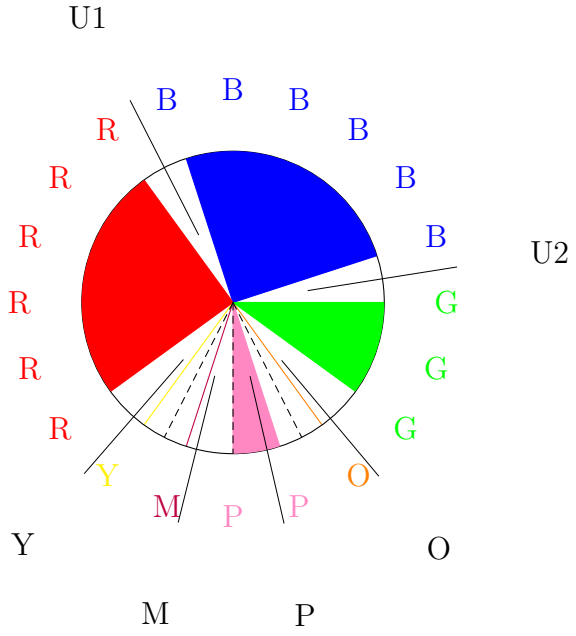

Figure 4.5: Probability wheel for Example 4.2.5

resulting maximum entropy distribution for the set of categories

$$\{\text{U1,U2,O,Y,M,P,G,R,B}\}$$

is

$$p_{maxE} = \{\frac{1}{20}, \frac{1}{20}, \frac{3}{40}, \frac{3}{40}, \frac{3}{40}, \frac{3}{40}, \frac{2}{20}, \frac{5}{20}, \frac{5}{20}\}.$$

◇

Example 4.2.6. Consider a multinomial data set with observed categories R, B, G, P and O and unobserved categories U1 and U2. The data are

$$(n_R, n_B, n_G, n_P, n_O) = (5, 5, 5, 5, 4).$$

The NPI-M algorithm initially assigns the lower probability $\frac{n_j-1}{n}$ to categories R, B, G, P and O. So $p(\mathrm{R}) = p(\mathrm{B}) = p(\mathrm{G}) = p(\mathrm{P}) = \frac{4}{24}$ and $p(\mathrm{O}) = \frac{3}{24}$. The algorithm initially assigns probability $\frac{1}{24}$ to the remaining categories.

We now reach the recursive part of the algorithm. For $i = 1$, $K(1) + K(2) = 0$. For $i = 2$, $K(2) + K(3) = 0$. For $i = 3$, $K(3) + K(4) = 1$ and $mass = 3$, so $K(i) + K(i+1) < mass$ and in this step of the algorithm $p(\mathrm{O})$ is increased from $\frac{3}{24}$ to $\frac{4}{24}$. For $i = 4$, $K(4) + K(5) = 5$ and $mass = 2$, therefore $K(i) + K(i+1) > mass$

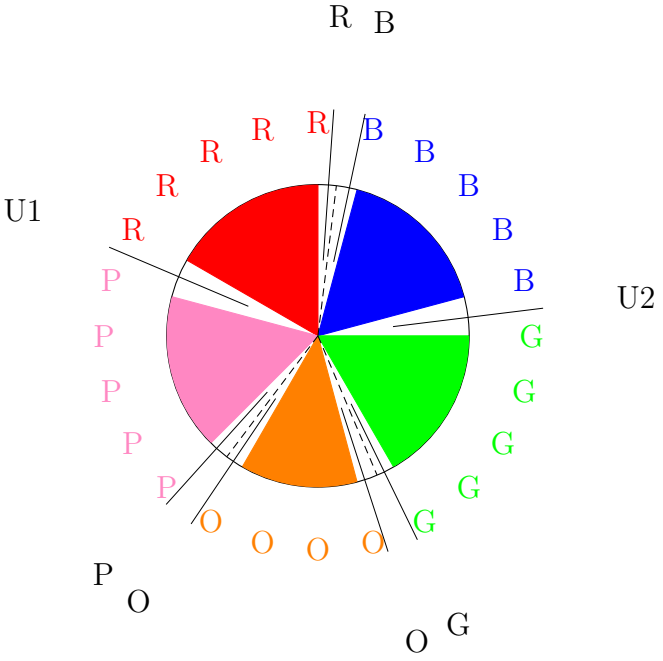

Figure 4.6: Probability wheel for Example 4.2.6

and

$$W = \min\{mass + 1 + K(4), K(4) + K(5)\} = mass + 1 + K(4) = 4.$$

This means that four of the five categories observed either four or five times are assigned the probability

$$p_{maxE}(c_j) = \frac{4}{24} + \frac{mass}{24(\min\{mass + 1 + K(4), K(4) + K(5)\})} = \frac{4}{24} + \frac{2}{4 \times 24} = \frac{3}{16},$$

while the remaining category is assigned the probability

$$p_{maxE}(c_j) = \frac{4}{24}.$$

Note that it does not matter which of the categories is assigned the smaller probability, since we sum over all categories when calculating the entropy. A possible configuration of the probability wheel is shown in Figure 4.6. The resulting maximum entropy distribution for the set of categories

$$\{\text{U1,U2,O,R,B,G,P}\}$$

is

$$p_{maxE} = \{\frac{1}{24}, \frac{1}{24}, \frac{4}{24}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}\}.$$

◇

4.3 Performance of the A-NPI-M algorithm

In order to assess the success of the A-NPI-M algorithm when used in building classification trees, experiments were carried out on forty data sets. These data sets were obtained from the UCI machine learning repository [4] and were chosen such that they vary greatly in terms of the sample size, the number and type of attribute variables and the number of categories. As a first step, an A-NPI-M classification tree was built for each data set. This was done using the tree-building process for interval probabilities described in Subsection 2.3.2, with the entropy values in the impurity measure (2.18) calculated using the A-NPI-M probabilities. Then, classification trees were produced for each data set using four alternative methods. The first of these alternative methods was the IDM method: the tree-building process remains the same, but the maximum entropy probabilities in the impurity measure (2.18) are based on the IDM with $s = 1$ (see Section 2.1) rather than the MNPI model. The remaining three alternative methods used all involve classical probabilities. Two of these methods use the ID3 [34] tree-building process but use varying impurity measures and the final method used was the C4.5 method [35], which is a more complex tree-building process that uses pruning to improve accuracy. The five classifiers resulting from these five methods were then compared in various ways.

The experiments were carried out using Weka software (see Subsection 2.3.3). First, the algorithms and methods required were implemented in Weka. Some preprocessing of the data was then carried out, using the filters in Weka. The missing value filter was applied, which replaces missing values with mean or modal values, and the discretise filter was applied, which discretises any continuous variables. Such preprocessing is necessary in order to apply our classification methods to the data in a straightforward way. However, it would be of interest to carry out further research on how NPI classification methods could be adapted to deal with data sets that do have missing values and on the effect of discretisation.

The classification trees were then built using Weka. A ten-fold cross-validation procedure was used (see Subsection 2.3.2) and this was repeated ten times for each data set. The comparisons between classifiers were based on the average numbers of correct classifications. In addition to a straightforward comparison based on which classifier had the highest number of correct classifications, various statistical tests were also used. A brief description of each of these tests is given in the following list. Further information on these tests can be found in [36].

1. Paired T-test: This test is used to compare two classifiers on a single data set. It checks whether one classifier is significantly better or worse than the other, by testing the null hypothesis that the difference between the numbers of correct classifications has a mean value of zero.

2. Wilcoxon signed-ranks test: This test is used to compare two classifiers on multiple data sets. For each data set, the test ranks the difference in performance of the two classifiers and these ranks are used to test the null hypothesis that there is no significant difference between the classifiers.

3. Friedman test: This test is used to compare multiple classifiers on multiple data sets. For each data set, the test ranks the classifiers and they are then compared in terms of their average rank. The null hypothesis is tested that there is no significant difference between these average ranks.

	IDM	A-NPI-M	IG	IGR	C4.5
IDM	-	(19/2/19)	(18/2/20)	(15/2/23)	(17/1/22)
NPI	(19/2/19)	-	(18/2/20)	(15/2/23)	(17/1/22)
IG	(20/2/18)	(20/2/18)	-	(18/3/19)	(19/1/20)
IGR	(23/2/15)	(23/2/15)	(19/3/18)	-	(21/1/18)
C4.5	(22/1/17)	(22/1/17)	(20/1/19)	(18/1/21)	-
W-L	15	15	-4	-20	-8

Table 4.1: Numbers of wins, ties and losses (W/T/L) based on average numbers of correct classifications

	IDM	A-NPI-M	IG	IGR	C4.5
IDM	-	(3/34/3)	(6/23/11)	(4/32/4)	(3/36/1)
NPI	(3/34/3)	-	(8/19/13)	(4/28/8)	(4/34/2)
IG	(11/23/6)	(13/19/8)	-	(6/30/4)	(11/21/8)
IGR	(4/32/4)	(8/28/4)	(4/30/6)	-	(7/24/9)
C4.5	(1/36/3)	(2/34/4)	(8/21/11)	(9/24/7)	-
W-L	3	7	-15	0	5

Table 4.2: Numbers of wins, ties and losses (W/T/L) based on the paired T-test carried out on numbers of correct classifications

$Classifier$	$Rank$
IDM	2.81
A-NPI-M	2.81
IG	3.02
IGR	3.25
C4.5	3.10

Table 4.3: Friedman ranks of the classifiers

The results of the experiments are shown in Tables 4.1 to 4.3. Table 4.1 shows the results of the straightforward comparison based on average numbers of correct classifications. Each column in the table corresponds to a particular classifier and shows the numbers of wins, ties and losses (W/T/L) for that classifier when compared pairwise with the other four classifiers. For example, (23/2/15) in the A-NPI-M column and IGR row means that the A-NPI-M classifier performed better than the IGR classifier on 23 data sets, equally well on 2 data sets and worse on 15 data sets. Table 4.2 shows the results of the paired T-test at a 5%

significance level. Again, each column shows the numbers of wins, ties and losses for a particular classifier. A win signifies that the performance of the classifier was significantly better at the 5% level, while a loss signifies that the performance was significantly worse. The Wilcoxon signed-ranks test showed that at a 5% significance level, the A-NPI-M classifier performs significantly better than the IGR classifier. The other pairwise comparisons did not highlight any significant differences. The Friedman ranks of the classifiers are shown in Table 4.3. The null hypothesis that the Friedman ranks are not significantly different is rejected and the Friedman ranks of the A-NPI-M classifier and the IDM classifier are found to be significantly higher than the ranks of the other classifiers at a 5% level of significance.

We see that the A-NPI-M and IDM classifiers have very similar performance, with the A-NPI-M classifier slightly outperforming the IDM classifier according to the paired T-test as shown in Table 4.2. Both perform better than the other classifiers considered here. Further detailed experiments on each data set individually, and also on further available data sets, would enable us to determine which classifier performs best on each data set and may give some insight into common characteristics of the data sets on which the A-NPI-M classifer performs well. This is not considered here, but further investigation would be useful for future research as it may allow us to give advice about which classification method would be best given specific features of a particular data set.

4.4 Comparison of A-NPI-M and NPI-M

In order to compare the A-NPI-M and NPI-M algorithms and to determine which of these is more successful when applied to classification trees, further experiments were carried out using Weka. The same forty data sets were used as in Section 4.3 and classification trees were built for each data set. When building the A-NPI-M and NPI-M classifiers, the tree-building procedure described in Subsection 2.3.2 was used and the entropy values in the impurity measure were computed using the A-NPI-M and NPI-M probabilities, respectively. Table 4.4 shows the percentages of

correct classifications achieved by these two classifiers for each of the forty data sets.

Dataset	NPI-M	A-NPI-M	Dataset	NPI-M	A-NPI-M
anneal	99.09	99.09	mfeat-morphological	69.78	69.78
arrhythmia	67.88	68.06	mfeat-pixel	79.99	79.92
audiology	85.04	85.04	mfeat-zernike	64.19	64.24
autos	78.45	78.25	nursery	95.15	94.99
balance-scale	69.59	69.59	optdigits	78.95	78.98
bridges-version1	67.74	67.74	page-blocks	96.08	96.10
bridges-version2	64.15	63.87	pendigits	89.37	89.37
car	90.13	90.13	postoperative-patient-data	71.11	71.11
cmc	48.98	48.98	primary-tumor	39.21	39.48
dermatology	93.43	93.46	segment	94.18	94.20
ecoli	80.19	80.19	soybean	93.29	93.35
flags	59.12	59.27	spectrometer	43.32	43.33
hypothyroid	99.33	99.33	splice	93.25	93.25
iris	93.40	93.40	sponge	94.48	94.48
letter	78.77	78.77	tae	46.78	46.78
lung-cancer	41.33	41.33	vehicle	69.39	69.39
lymphography	73.68	73.68	vowel	75.92	75.95
mfeat-factors	81.71	81.68	waveform	73.99	73.99
mfeat-fourier	68.90	68.92	wine	92.02	92.02
mfeat-karhunen	73.14	73.15	zoo	95.53	95.53

Table 4.4: Percentages of correct classifications

The A-NPI-M and NPI-M classifiers were compared with each other and with the IDM classifier using the paired T-test, the Wilcoxon signed-ranks test and the Friedman test, all of which are described in Section 4.3. A straightforward comparison was also carried out based on which classifier had the highest number of correct classifications. Table 4.5 shows the results of the comparison based on average numbers of correct classifications and Table 4.6 shows the results of the paired T-test at a 5% significance level. The Wilcoxon signed-ranks test showed no significant differences between any of the classifiers. The Friedman ranks of the classifiers, shown in Table 4.7, are not significantly different at a 5% level of significance.

	IDM	A-NPI-M	NPI-M
IDM	-	(21/2/17)	(21/2/17)
A-NPI-M	(17/2/21)	-	(7/20/13)
NPI-M	(17/2/21)	(13/20/7)	-
W-L	-8	10	-2

Table 4.5: Numbers of wins, ties and losses (W/T/L) based on average numbers of correct classifications

	IDM	A-NPI-M	NPI-M
IDM	-	(3/34/3)	(3/34/3)
A-NPI-M	(3/34/3)	-	(1/39/0)
NPI-M	(3/34/3)	(0/39/1)	-
W-L	0	-1	1

Table 4.6: Numbers of wins, ties and losses (W/T/L) based on the paired T-test carried out on numbers of correct classifications

Classifier	*Rank*
IDM	2.10
A-NPI-M	1.85
NPI-M	2.05

Table 4.7: Friedman ranks of the classifiers

We see from the results in Tables 4.5, 4.6 and 4.7 that all three classifiers are quite similar, with the NPI-based classifiers slightly outperforming the IDM classifier. The A-NPI-M and NPI-M classifiers perform in a very similar way. Table 4.5 shows that the A-NPI-M classifier gives more correct classifications than the NPI-M classifier on 13 data sets; however, for most data sets the difference in numbers of correct classifications is negligible and we see from Table 4.6 that there is only one data set on which the two classifiers differ significantly. On this data set, entitled the nursery data set, the paired T-test shows us that the NPI-M classifier gives a significantly higher number of correct classifications.

Weka was used to analyse the nursery data set in more detail, in order to give some insight about any characteristics of this data set that may cause the significant difference in the performance of the A-NPI-M and NPI-M classifiers. The nursery data set is taken from a set of applications to a private nursery school, and contains a total of 12960 observations. There are eight attribute variables and five categories c_1 to c_5 where an observation is classified in terms of how suitable the applicant is for acceptance at the school. The classification trees built using Weka show that the health of the applicant is the most informative attribute

variable, so this is used for splitting at the root node (see Figure 4.7). There are

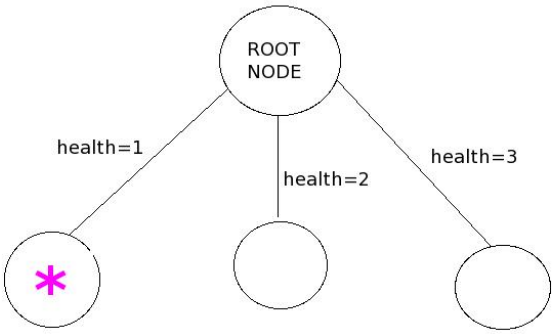

Figure 4.7: Diagram showing first split in classification tree for nursery data set

three possible values of this attribute variable, labelled 1 to 3, and in the branch health=1 (i.e. at the node marked * in Figure 4.7), the observation counts for the set of categories $\{c_1, ..., c_5\}$ are $\{0, 0, 0, 1854, 2466\}$. All observations in categories c_1 to c_3 have been eliminated by this split, showing that health is a very informative attribute variable. At this point, the A-NPI-M algorithm returns the distribution $\{\frac{1}{6480}, \frac{1}{6480}, \frac{1}{6480}, \frac{1853}{4320}, \frac{2465}{4320}\}$ corresponding to categories $c_1, ..., c_5$, whereas the NPI-M algorithm returns the distribution $\{\frac{1}{8640}, \frac{1}{8640}, \frac{1}{4320}, \frac{1853}{4320}, \frac{2465}{4320}\}$. The fact that this discrepancy occurs so high up in the tree may be a reason for the significantly different performance of the A-NPI-M and NPI-M classifiers on this data set, but a more detailed investigation would be needed in order to gain a full understanding of this difference.

It should also be noted here that, for the nursery data set, there is a natural ordering of the categories. The MNPI model explicitly assumes no specific ordering of the categories; however, NPI for ordinal data is currently under development and its application to classification problems is an interesting and important topic for future research.

Further analysis of each of the forty data sets, and also of further available data sets, would allow us to discover common characteristics of the data sets on which the NPI-M classifier performed best. This is not considered here, but it

would be useful for future research as it may enable us to make an informed decision about which classification method to use given a particular data set.

4.5 Split variable selection bias

In Sections 4.1 to 4.4, we considered classification trees that employ an estimator of the Shannon entropy H (see (2.13)) as the impurity measure used to select split variables. This estimator, $\widehat{H}$ (see (2.14)), is widely and successfully used in classification. However, $\widehat{H}$ is a biased estimator of the true Shannon entropy. The bias is linked to the numbers of possible values of the attribute variables and means that when selecting variables for splitting, the tree-building method tends to favour attribute variables with high numbers of distinct values over equally informative attribute variables with fewer values. A study by Strobl [39] showed that this effect is particularly pronounced when the attribute variables in question are not very informative.

The theoretical Shannon entropy, H, is calculated using the true category probabilities $p(c_j)$, whereas $\widehat{H}$ is a function of some estimate $\widehat{p}$ of this distribution. A naive approach to estimating the probabilities $p(c_j)$ is to use the observed relative frequencies. The estimated Shannon entropy is then calculated via (2.14) together with $\widehat{p}(c_j) = \frac{n_j}{n}$. The bias of this estimate was computed analytically by Miller [33]. It was shown that the expected value of the estimator was equal to

$$E(\widehat{H}(\widehat{p})) = E(-\sum_{j=1}^{K} \frac{n_j}{n} \log[\frac{n_j}{n}]) = H(p) - \frac{K-1}{2n} + O(n^{-2}). \tag{4.6}$$

Miller proposed that the terms of order n^{-2} should be ignored, due to the fact that these terms involve the unknown probabilities $p(c_j)$. He suggested the following correction to give a less biased estimator of the Shannon entropy:

$$\widehat{H}_{Miller}(\widehat{p}) = \widehat{H}(\widehat{p}) + \frac{K-1}{2n}. \tag{4.7}$$

As discussed in Chapter 2, when using interval probabilities the distribution used to estimate the true category probabilities is the maximum entropy distribution. The

probabilities $p_{maxE}(c_j)$ that are attained depend on the interval probability model being used. Strobl [39] investigated split variable selection bias in classification trees constructed using the IDM [41]. In this method, the maximum entropy probability for each category c_j is derived from the IDM interval probabilities $[\underline{P}_{IDM}(Y_{n+1} \in c_j), \overline{P}_{IDM}(Y_{n+1} \in c_j)] = [\frac{n_j}{n+s}, \frac{n_j+s}{n+s}]$. Based on the expected value of $\widehat{H}$ derived by Miller, it was suggested by Strobl that the entropy estimator $\widehat{H}$ be replaced by

$$\widehat{H}_{Miller}(\widehat{p}) = \widehat{H}(\widehat{p}) + \frac{K-1}{2(n+s)}$$

when using the IDM. This correction approach was tested in a simulation study [39], and the corrected estimator was found to perform better than the original with regard to split variable selection.

We now discuss the use of a bias correction when building classification trees using NPI. As explained in Section 4.2, the maximum entropy distribution within the NPI structure is achieved by the NPI-M algorithm. Consider some node $(X_i = x_i)$ which is generated by splitting on the attribute variable X_i. Let $n^{(X_i=x_i)}$ represent the total number of observations at this node and let $n_j^{(X_i=x_i)}$ represent the number of observations in category c_j at this node. At the node $(X_i = x_i)$, the maximum entropy probability for each category, $p_{maxE}^{(X_i=x_i)}(c_j)$, will always lie within the interval $[\frac{n_j^{(X_i=x_i)}-1}{n^{(X_i=x_i)}}, \frac{n_j^{(X_i=x_i)}+1}{n^{(X_i=x_i)}}]$. The standard entropy estimator, used in Sections 4.3 and 4.4, is

$$\widehat{H}(p_{maxE}^{(X_i=x_i)}) = -\sum_{j=1}^{K} p_{maxE}^{(X_i=x_i)}(c_j) \log[p_{maxE}^{(X_i=x_i)}(c_j)]. \tag{4.8}$$

Based on Miller's derivation of the extent of the bias in $\widehat{H}$ (see (4.6)), it is reasonable to assume that the bias in (4.8) is equivalent. We therefore replace this estimator by

$$\widehat{H}_{Miller}(p_{maxE}^{(X_i=x_i)}) = \widehat{H}(p_{maxE}^{(X_i=x_i)}) + \frac{K-1}{2n^{(X_i=x_i)}}. \tag{4.9}$$

In order to test this bias correction, experiments in Weka were carried out using the forty data sets introduced in Section 4.3. The NPI-M algorithm was implemented to build classification trees for each data set, first with the impurity measure calculated using the standard entropy estimator (4.8) and secondly with the

impurity measure calculated using the corrected entropy estimator (4.9). The two methods were compared based on which gave the highest number of correct classifications. Further comparisons were carried out using the paired T-test with a 5% significance level, the Wilcoxon signed-ranks test and the Friedman test. These tests are described in Section 4.3.

With regard to the straightforward comparison based on numbers of correct classifications, the numbers of wins, ties and losses (W/T/L) for the bias-corrected method against the original method were (6/3/31), i.e. the bias-corrected method performed better than the original method on 6 data sets but worse on 31 data sets. The equivalent results of the paired T-test were (1/23/16), i.e. the performance of the bias-corrected method was significantly better on only 1 data set but was significantly worse on 16 data sets. The Wilcoxon signed-ranks test showed that at a 5% significance level, the original method performed better than the bias-corrected method. The Friedman ranks of the bias-corrected and original methods were shown to be 1.8125 and 1.1875 respectively. These are found to be significantly different at a 5% level of significance.

We see from the results of these experiments that when building classification trees using NPI, introducing the above bias correction to $\widehat{H}$ does not appear to improve classifier performance. The original method appears to be more successful, so we cannot recommend at this stage that the method should be adapted. We therefore do not consider further the issue of split variable selection bias here; however, further exploration of this is an important topic for future research and a more detailed study is required in order to come to a final judgement on whether this or any other bias correction would be useful.

4.6 Concluding remarks

In this chapter we presented applications of NPI to classification. We focused on classification trees, since this type of classification does not make explicit use of

parameters in the way that other types such as naive classification do (see Section 2.3). However, it would be of interest to investigate applications of NPI to other types of classification in the future.

In this chapter, the use of the MNPI model for building classification trees was considered. Two algorithms were presented: the A-NPI-M algorithm for finding the maximum entropy distribution within the credal set of distributions associated with the NPI lower and upper probabilities and the NPI-M algorithm for finding the maximum entropy distribution within the NPI structure $\mathcal{M}_{NPI}$. These algorithms were used to build classification trees for forty data sets and experiments were carried out to measure and compare the performance of the algorithms. Finally, a bias correction to the Shannon entropy estimator was investigated, as a possible variation on the tree-building method. It would be of interest to extend this research further and to investigate more fully the use of a bias correction, as although we did not find this correction to be useful for our purposes, it has proved to be successful in a number of studies. It would also be interesting to consider the use of other impurity measures when building classification trees using NPI. A more detailed analysis of a wider range of data sets would be beneficial, including an investigation of common characteristics of the data sets on which the NPI-based classifiers perform well, as this may allow us to identify future data sets for which a NPI-based classifier would be a suitable choice.

Another important extension to this work is the development of imprecise classification trees, which may sometimes return a set of categories rather than a single most probable category. Imprecise classification has been presented in the literature [46,48], as mentioned in Subsection 2.3.1, and it seems natural to consider imprecise classification using NPI.

We return to the subject of classification in Section 5.4, where we present an algorithm for approximating the maximum entropy distribution consistent with the NPI model for data described at subcategory level as well as at main category level.

Chapter 5

NPI for subcategory data

In this chapter we present an extension of nonparametric predictive inference for multinomial data such that subcategories may be included in our inferences as well as main categories. This is motivated by the hierarchical structure that is inherent in some multinomial data sets. As before, there is no ordering of the main categories, and we also assume that for a single main category there is no ordering of its subcategories. We still use the probability wheel representation that was explained in Chapter 2, but we now include lines and subsegments to represent these subcategories. As in the original MNPI model, slices are assigned an area of $\frac{1}{n}$, representing the probability $\frac{1}{n}$ for the event that a future observation falls in any given slice. In addition to the assumptions underlying the MNPI model, the extended model presented here requires that two or more lines representing the same subcategory are always positioned next to each other on the wheel and that different subcategories within the same main category are always grouped together in one single segment of the wheel. Also, if a slice is bordered by two lines representing the same subcategory, it must be assigned to this subcategory. Multiple slices assigned to the same subcategory must always be grouped together in one single subsegment. We henceforth refer to this model as the Sub-MNPI model.

As in the original MNPI model (see Section 2.2), we assume that there are K main categories in total and that the data consist of n_j observations in main category c_j, for $j = 1, ..., K$. We have observed k main categories and the remaining

$K - k$ main categories are unobserved. We also assume that some observations may belong to a particular subcategory. Subcategories are denoted by s_{j,i_j}, where $s_{j,i_j} \subseteq c_j$. Suppose that there is a total of K_j subcategories in main category c_j and that we have observed k_j of these subcategories. Let n_{j,i_j} be the total number of observations in subcategory s_{j,i_j}. Some main categories may not consist of subcategories, or may only be described at main category level, in which case we continue to denote these simply by c_j. Such categories are referred to as main-only categories, distinct from main categories which may or may not have specified subcategories.

As is the case for the original MNPI model, it may be that we know the total numbers of possible main categories and subcategories, or it may be that these quantities are unknown. In Section 5.1, we consider the case where the quantities K and K_j, $j = 1, ..., K$, are known. We define the general event of interest and derive the NPI lower and upper probabilities for this event. In Section 5.2, we formulate some general properties of the NPI lower and upper probabilities presented in Section 5.1. In Section 5.3, we consider the case where K and K_j, $j = 1, ..., K$, are unknown. In this case there are two general events of interest. We define these events and derive the NPI lower and upper probabilities for both events. Finally, in Section 5.4, we revisit the application of NPI to classification trees discussed in Chapter 4 and we propose an algorithm for approximating the maximum entropy distribution consistent with the Sub-MNPI model that is presented in Section 5.1. The inferences throughout this chapter are restricted to a single future observation; however, it will be interesting to consider more general inferences about multiple future observations as an extension to this work.

5.1 Known numbers of (sub)categories

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

$$Y_{n+1} \in \bigcup_{j \in J} c_j \cup \bigcup_{j \in J^*} \bigcup_{i_j \in I_j} s_{j,i_j} \tag{5.1}$$

where $J \cap J^* = \emptyset$, $J \subseteq \{1, ..., K\}$, $J^* \subseteq \{1, ..., K\}$ and $I_j \subseteq \{1, ..., K_j\}$ for $j = 1, ..., K$. We also define $\overline{I_j} = \{1, ..., K_j\} \backslash I_j$. This notation allows us to describe events which contain only specific subcategories of particular main categories, whilst also retaining the possibility of considering some main categories as a whole. We refer to the general event (5.1) as E. Note that in examples, E denotes specific events of the form shown in (5.1).

As in Section 2.2, k categories have been observed and these are labelled $c_1, ..., c_k$. As in Subsection 2.2.1, we have $OJ = J \cap \{1, ..., k\}$, which is the index-set of observed main-only categories in E, and we define $\mid OJ \mid = r_{main}$. Also, recall that $UJ = J \cap \{k + 1, ..., K\}$, which is the index-set of unobserved main-only categories in E, and let $\mid UJ \mid = l_{main}$. Similarly, we define

$$OJ^* = J^* \cap \{1, ..., k\},$$

where $\mid OJ^* \mid = r_{sub}$. OJ^* is the index-set of observed main categories in E which are described at subcategory level. We also define

$$UJ^* = J^* \cap \{k + 1, ..., K\},$$

where $\mid UJ^* \mid = l_{sub}$. UJ^* is the index-set of unobserved main categories in E which are described at subcategory level. Let $r = r_{main} + r_{sub}$, and let $l = l_{main} + l_{sub}$.

Suppose that k_j subcategories of main category c_j have been observed and are labelled $s_{j,1}, ..., s_{j,k_j}$. Let

$$OI_j = I_j \cap \{1, ..., k_j\},$$

where $\mid OI_j \mid = r_j$, for $j = 1, ..., K$. OI_j is the index-set of observed subcategories in E. Also suppose that

$$UI_j = I_j \cap \{k_j + 1, ..., K_j\},$$

where $\mid UI_j \mid = l_j$, for $j = 1, ..., K$. UI_j is the index-set of unobserved subcategories in E.

Let $\overline{OI_j} = \overline{I_j} \cap \{1, ..., k_j\}$, where $\mid \overline{OI_j} \mid = \overline{r_j}$, and let $\overline{UI_j} = \overline{I_j} \cap \{k_j + 1, ..., K_j\}$, where $\mid \overline{UI_j} \mid = \overline{l_j}$.

We consider the NPI lower and upper probabilities for E (5.1).

5.1.1 Lower probability

First, we consider the NPI lower probability for the event E (5.1). This is found by minimising the number of slices of the probability wheel which must be assigned to a main category or subcategory in E.

There are four different cases which may arise, depending on the numbers of categories and subcategories which have already been observed, and these are considered separately.

Case 1: $\boldsymbol{r \leq K - r - l}$ and $\boldsymbol{r_j - 1 \leq K_j - r_j - l_j}$ for all $\boldsymbol{j}$

When $r \leq K - r - l$, the number of main categories not in E is greater than or equal to the number of observed main categories in E. This means that all separating slices on the wheel between different observed main categories in E can be assigned to a main category which is not in E. Similarly, when $r_j - 1 \leq K_j - r_j - l_j$, all separating slices between different observed subcategories in E can be assigned to subcategories of main category c_j which are not in E. Therefore, the only slices which we are forced to assign to E are those which lie between lines representing the same main-only category in E and those which lie between lines representing the same subcategory in E. This leads to

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \sum_{j \in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n}. \tag{5.2}$$

Example 5.1.1. Consider a multinomial data set with possible main categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). These main categories are labelled 1 to 6 respectively. Observations in B are further classified as light blue (LB), medium blue (MB), dark blue (DB) or other blue (OB) and observations in

G are further classified as light green (LG), dark green (DG) or other green (OG). The data set consists of eight observations altogether, including 1 LB, 1 MB, 2 DB, 1 LG, 1 DG, 1 R and 1 Y.

Suppose that we are interested in the event $Y_9 \in \{\text{LB,DB,LG,P}\}$. Then $K = 6$, $r = 2$ and $l = 1$. For main categories described at subcategory level, the values of K_j, r_j and l_j are shown in Table 5.1. This example illustrates the

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

Table 5.1: Values of K_j, r_j and l_j for Example 5.1.1

situation where $r \leq K - r - l$ and $r_j - 1 \leq K_j - r_j - l_j$ for all j. We can therefore find a configuration of the probability wheel such that all main categories in E are separated by main categories not in E and within each segment all subcategories in E are separated by subcategories not in E. Figure 5.1 shows one such configuration,

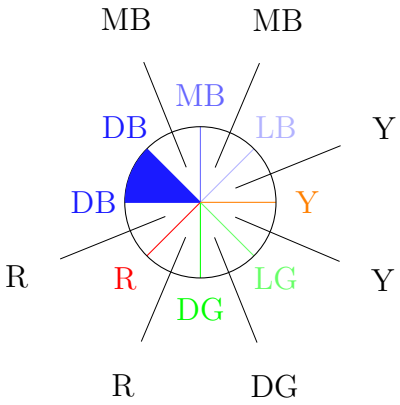

Figure 5.1: Probability wheel for Example 5.1.1

where B and G are separated by R and Y, and LB and DB are separated by MB. The only part of the wheel which must be assigned to a category in E is therefore the slice between the two DB observations. This means that the NPI lower probability for the event E is $\frac{1}{8}$. This also follows from (5.2): the set OJ is empty (since P is not yet observed), the set OJ^* contains B and G, the set OI_1

contains LB and DB and the set OI_2 contains LG, therefore (5.2) gives

$$\underline{P}(E) = 0 + \frac{1-1}{8} + \frac{2-1}{8} + \frac{1-1}{8} = \frac{1}{8}.$$

◇

Case 2: $r > K - r - l$ and $r_j - 1 \leq K_j - r_j - l_j$ for all j

As in Case 1, all lines representing different observed subcategories in E can be separated by subcategories which are not in E. Now, though, the number of main categories not in E is smaller than the number of observed main categories in E. This means that not all of the observed main categories in E can be separated by main categories not in E. There are $r - (K - r - l) = 2r + l - K$ separating slices between main categories which cannot be filled in this way. However, if we have subcategories which are not in E but which are part of a main category that appears in E, it may be possible to utilise these subcategories to separate observed main categories in E. In order to see how this could work, we need to consider the configuration of the slices carefully.

Recall that $\overline{r_j} + \overline{l_j}$ is the total number of subcategories within category j that are not in E. Within each segment of the wheel, if $\overline{r_j} + \overline{l_j} > r_j$ then we can find a configuration such that all observations of subcategories in E are separated and also each end of the segment (i.e. the line on either end of the segment) represents a subcategory not in E. If $\overline{r_j} + \overline{l_j} = r_j$ then the best configuration of the segment is such that one end of the segment represents a subcategory in E and one end represents a subcategory not in E. Finally, if $\overline{r_j} + \overline{l_j} < r_j$, we cannot find a configuration such that the ends of the segment represent a subcategory not in E whilst still satisfying the requirement that all observations of subcategories in E are separated. So, within each main category such that $j \in J^*$, $\overline{r_j} + \overline{l_j} - (r_j - 1) = \overline{r_j} + \overline{l_j} - r_j + 1$ subcategories would potentially be available for separating main categories in E.

It is important to remember that the subcategories of a single main category

must always be grouped together in a single segment of the wheel. This means that the maximum number of subcategories which could be assigned to separating slices between main categories would be two per observed main category and one per unobserved main category. The number of separating slices which can potentially be filled in this way is therefore

$$S_M = \sum_{j \in OJ^*} \min\{(\overline{r_j} + \overline{l_j} - r_j + 1)^+, 2\} + \sum_{j \in UJ^*} \min\{\overline{l_j}, 1\} \tag{5.3}$$

where the notation x^+ represents $\max\{x, 0\}$. So, the overall number of slices which still cannot be filled by a main category or subcategory not in E is $2r + l - K - S_M$, provided that this is a number greater than or equal to zero. This leads to

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \sum_{j \in OJ^*} \sum_{i \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{1}{n}(2r + l - K - S_M)^+. \tag{5.4}$$

Example 5.1.2. Consider the data set described in Example 5.1.1. Suppose that we are interested in the event $Y_9 \in \{\text{LB,DB,LG,R,Y,P}\}$. We have $K = 6$, $r = 4$ and $l = 1$. For main categories described at subcategory level, the values of K_j, r_j and l_j are shown in Table 5.2. We have the situation where $r_j - 1 \leq K_j - r_j - l_j$

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

Table 5.2: Values of K_j, r_j and l_j for Example 5.1.2

for all j, but $r > K - r - l$. We therefore cannot find any configuration of the probability wheel such that all observed main categories in E are separated by main categories not in E. However, we can still separate subcategories in E within each individual segment. In this example, $2r + l - K = 3$ and $S_M = 3$. So whilst we cannot separate all observed main categories in E using other main categories, we can in fact use subcategories which are not in E but which are part of a main category that appears in E.

In this example, the main category B appears in E, but the subcategories MB and OB are not included in E. Similarly, the main category G appears in E, but the subcategories DG and OG are not included in E. Figure 5.2 shows one

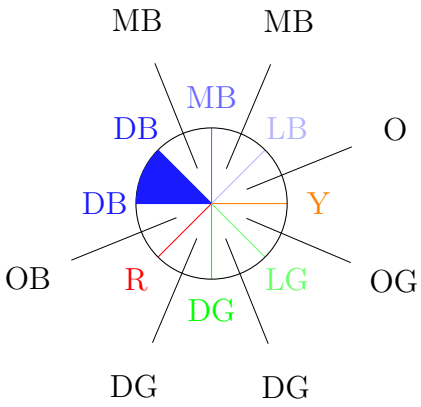

Figure 5.2: Probability wheel for Example 5.1.2

possible configuration where all elements of E are separated and illustrates the case where O separates B and Y, OB separates R and B, DG separates G and R, and OG separates G and Y. The only part of the wheel which must be assigned to E is the slice between the two DB observations, leading to a NPI lower probability of $\frac{1}{8}$ for the event E. This lower probability can be verified using (5.4). The set OJ contains R and Y, the set OJ^* contains B and G, the set OI_1 contains LB and DB and the set OI_2 contains LG. Also, $2r+l-K=3$ and $S_M=3$, therefore (5.4) gives

$$\underline{P}(E)=\frac{1-1}{8}+\frac{1-1}{8}+\frac{1-1}{8}+\frac{2-1}{8}+\frac{1-1}{8}+0=\frac{1}{8}.$$

◇

Case 3: $r \leq K-r-l$ and $r_j-1>K_j-r_j-l_j$ for at least one j

In this situation, we can separate all lines corresponding to different observed main categories in E by main categories not in E. This is because the number of main categories not in E is greater than or equal to the number of observed main categories in E. However, within main categories for which $r_j-1>K_j-r_j-l_j$ holds, we cannot separate all the observed subcategories in E, since the number of subcategories not in E is smaller than the number of separating slices between observed subcategories

in E. For such a category, there are $r_j - 1 - (K_j - r_j - l_j) = 2r_j + l_j - K_j - 1$ of these separating slices remaining and these have to be assigned to E. This leads to

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \sum_{j \in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{1}{n} \sum_{j \in OJ^*} (2r_j + l_j - K_j - 1)^+. \quad (5.5)$$

Example 5.1.3. Consider the data set described in Example 5.1.1. Suppose that we are interested in the event $Y_9 \in \{\text{LB,MB,DB,LG,P}\}$. We have $K = 6$, $r = 2$ and $l = 1$. For main categories which have subcategories, the values of K_j, r_j and l_j are given in Table 5.3. So we have the situation where $r \leq K - r - l$, but

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

Table 5.3: Values of K_j, r_j and l_j for Example 5.1.3

$r_j - 1 > K_j - r_j - l_j$ for the case $j = 1$, i.e. for main category B. We can therefore separate all observed main categories in E by main categories not in E, but within the B segment we are unable to find a configuration such that all subcategories in E are separated. Figure 5.3 illustrates this and shows a configuration where B and

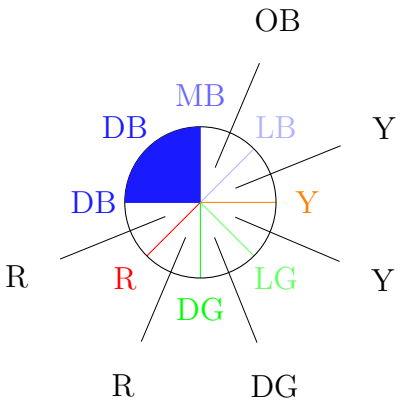

Figure 5.3: Probability wheel for Example 5.1.3

G are separated by R and Y. Looking specifically at the B segment, we see that although OB can be used to separate LB and MB, the slice between MB and DB

then has to belong to either MB or DB and we are therefore forced to assign this slice to E. This means that the NPI lower probability for the event E is $\frac{2}{8}$. We can verify this using (5.5). We see that the set OJ is empty, the set OJ^* contains B and G, the set OI_1 contains LB, MB and DB and the set OI_2 contains LG. Also, $\sum_{j\in OJ^*} \max\{2r_j + l_j - K_j - 1, 0\} = 1$, therefore (5.5) gives

$$\underline{P}(E) = 0 + \frac{1-1}{8} + \frac{2-1}{8} + \frac{1-1}{8} + \frac{1-1}{8} + \frac{1}{8} = \frac{2}{8}. \qquad \diamond$$

Case 4: $r > K - r - l$ and $r_j - 1 > K_j - r_j - l_j$ for at least one j

In this situation, not all of the observed main categories in E can be separated by main categories not in E. As explained in Case 2, there are $2r + l - K$ separating slices between main categories which cannot be filled in this way. However, as illustrated in Example 5.1.2, we may still be able to separate observed main categories in E using subcategories which are not in E but which are within a main category that appears in E. As shown in Case 2, the number of subcategories which can potentially be used to separate main categories in E is S_M as given by (5.3).

In addition, due to the fact that $r_j - 1 > K_j - r_j - l_j$ for some j, some segments have observed subcategories in E that cannot be separated by subcategories not in E. As explained in Case 3, this means that there are $2r_j + l_j - K_j - 1$ separating slices remaining, which we are forced to assign to E. This leads to

$$\begin{aligned}\underline{P}(E) = &\sum_{j\in OJ} \frac{n_j - 1}{n} + \sum_{j\in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{1}{n}(2r + l - K - S_M)^+ \\ &+ \frac{1}{n} \sum_{j\in OJ^*} (2r_j + l_j - K_j - 1)^+.\end{aligned} \tag{5.6}$$

Example 5.1.4. Consider the data set described in Example 5.1.1. Suppose that we are interested in the event $Y_9 \in$ {LB,MB,DB,LG,R,Y,P}. We have $K = 6$, $r = 4$ and $l = 1$. For main categories described at subcategory level, the values of K_j, r_j and l_j are shown in Table 5.4. Here we have the situation where $r_j - 1 > K_j - r_j - l_j$

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

Table 5.4: Values of K_j, r_j and l_j for Example 5.1.4

for $j = 1$, i.e. for main category B, and also $r > K - r - l$. We are therefore unable to separate all subcategories in E within the B segment. Furthermore, we are unable to configure the probability wheel such that all observed main categories in E are separated by main categories not in E. In this example, $2r + l - K = 3$ and $S_M = 2$. So whilst we can use some subcategories which are not in E but which are part of a main category that appears in E, there is still one separating slice between main categories which has to be assigned to E. Figure 5.4 shows a possible

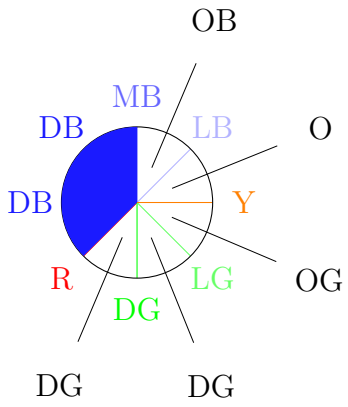

Figure 5.4: Probability wheel for Example 5.1.4

configuration of the wheel such that O separates B and Y, OG separates Y and G, and DG separates G and R. There is then no way of separating R and B by a main category or subcategory not in E and we are therefore forced to assign this slice to E. Looking specifically at the B segment, we see that OB separates LB and MB but the slice between MB and DB then has to be assigned to E. This leads to a NPI lower probability of $\frac{3}{8}$ for the event E. This lower probability can be verified using (5.6). We see that the set OJ contains R and Y, the set OJ^* contains B and G, the set OI_1 contains LB, MB and DB and the set OI_2 contains LG. Also, $2r+l-K = 3$,

$S_M = 2$ and $\sum_{j \in OJ^*} \max\{2r_j + l_j - K_j - 1, 0\} = 1$, therefore (5.6) gives

$$\underline{P}(E) = \frac{1-1}{8} + \frac{1-1}{8} + \frac{1-1}{8} + \frac{2-1}{8} + \frac{1-1}{8} + \frac{1-1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}.$$

$\diamond$

The NPI lower probability formulae for the four cases in this subsection can be combined to give the following general expression:

$$\begin{aligned}\underline{P}(E) = &\sum_{j \in OJ} \frac{n_j - 1}{n} + \sum_{j \in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{1}{n}(2r + l - K - S_M)^+ \\ &+ \frac{1}{n} \sum_{j \in OJ^*} (2r_j + l_j - K_j - 1)^+.\end{aligned} \tag{5.7}$$

5.1.2 Upper probability

We now derive the NPI upper probability for the event E (5.1). This is found by constructing a configuration of the probability wheel which maximises the number of slices assigned to E. We do this by considering which slices can definitely not be assigned to E and are accounted for by the $k - r$ observed main categories not in E or by the $\overline{r_j}$ observed subcategories not in E. In order to construct such a configuration, we first need to think about the various ways in which we can separate lines or segments on the wheel representing different main categories which either are not in E or are present in E but have neither end of their segment in E.

First, we can separate these main categories using unobserved main categories in E. There are l of these categories. Secondly, we can separate using observed main-only categories in E. There are r_{main} such categories. Finally, we can separate using the other observed main categories in E, provided that the configuration of the relevant segment is such that each end represents a subcategory in E. There are r_{sub} main categories in E that are described at subcategory level. For a segment to have the required configuration, the category must satisfy $k_j - r_j + 1 \leq r_j + l_j$. This is because we need $k_j - r_j - 1$ subcategories in E to ensure that all subcategories not in E are separated and a further two to ensure that both ends of the segment are in E. We define the number of main categories which satisfy this condition as $\tilde{r}_{sub}$. We define

the number of main categories which are present in E but have neither end of their segment belonging to E, i.e. the number which satisfy $k_j - r_j - 1 \geq r_j + l_j$, as r^0_{sub}. We define the number of main categories which are present in E but have only one end of their segment belonging to E, i.e. the number which satisfy $k_j - r_j = r_j + l_j$, as r^1_{sub}.

As in Subsection 5.1.1, there are four different cases which we consider individually.

Case 1: $(k-r) + r^0_{sub} \leq l + r_{main} + \tilde{r}_{sub}$, $k_j - r_j - 1 \leq r_j + l_j$ for all j

When $(k-r) + r^0_{sub} \leq l + r_{main} + \tilde{r}_{sub}$, the number of main categories in E which are either main-only or have both ends of their segment in E is greater than or equal to the number of observed main categories which are either not in E at all or are present in E but have neither end of their segment in E. All k separating slices between two lines representing different main categories can therefore be assigned to E.

Similarly, when $k_j - r_j - 1 \leq r_j + l_j$ within main category c_j, there are enough subcategories in E within each single segment to separate all the different observed subcategories not in E. Therefore all $k_j - 1$ separating slices between two lines representing different subcategories can be assigned to E.

In addition, we must assign to E the slices between lines representing the same main-only category in E and the slices between lines representing the same subcategory in E. This leads to

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{k}{n} + \sum_{j \in OJ^*} \Big(\sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{k_j - 1}{n} \Big). \tag{5.8}$$

Example 5.1.5. Consider the data set used in Examples 5.1.1 to 5.1.4. This is a multinomial data set with possible main categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). These main categories are labelled 1 to 6 respectively. Observations in B are further classified as light blue (LB), medium blue (MB), dark blue (DB) or other blue (OB) and observations in G are further classified as light green (LG), dark green (DG) or other green (OG). The data set

consists of eight observations altogether, including 1 LB, 1 MB, 2 DB, 1 LG, 1 DG, 1 R and 1 Y. Suppose that we are interested in the event $Y_9 \in$ {LB,DB,LG,P}. Then $k = 4$, $r_{main} = 0$, $r_{sub} = 2$, $r = 2$ and $l = 1$. For main categories which have subcategories, the values of k_j, r_j and l_j are shown in Table 5.5. The values in

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

Table 5.5: Values of k_j, r_j and l_j for Example 5.1.5

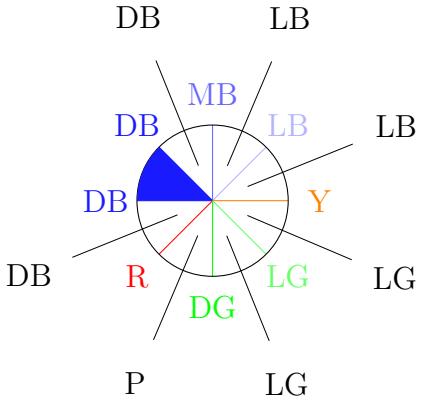

Figure 5.5: Probability wheel for Example 5.1.5

Table 5.5 show that the condition $k_j - r_j + 1 \leq r_j + l_j$ is satisfied by main category B but not by main category G, so we have $\tilde{r}_{sub} = 1$. We also have $r^0_{sub} = 0$. This example illustrates the situation where $(k - r) + r^0_{sub} \leq l + r_{main} + \tilde{r}_{sub}$ and also $k_j - r_j - 1 \leq r_j + l_j$ for all j. We can find a configuration of the probability wheel such that every slice is accounted for by a main category or subcategory in E. Figure 5.5 shows such a configuration with R and Y separated by B, R and G separated by P, and G and Y separated by LG (since LG is in E). This leads to a NPI upper probability of 1 for the event E. This upper probability can be verified using (5.8): we see that the set OJ is empty, the set OJ^* contains B and G, the set OI_1 contains LB and DB and the set OI_2 contains LG, therefore (5.8) gives

$$\overline{P}(E) = \frac{4}{8} + \frac{1-1}{8} + \frac{2-1}{8} + \frac{2}{8} + \frac{1-1}{8} + \frac{1}{8} = 1.$$

$\diamond$

Case 2: $(k-r)+r^0_{sub} > l + r_{main} + \tilde{r}_{sub}$, $k_j - r_j - 1 \leq r_j + l_j$ for all j

As in Case 1, within each individual segment there are enough subcategories in E to separate all observed subcategories not in E. We therefore find that within each segment, all $k_j - 1$ separating slices between two lines representing different subcategories can be assigned to E.

However, the number of observed main categories which have neither end of their segment in E is greater than the number of main categories in E which are either main-only or have both ends of their segment in E. There are $(k-r)+r^0_{sub} - (l + r_{main} + \tilde{r}_{sub})$ separating slices which we cannot assign to E, in other words only $k - ((k-r)+r^0_{sub} - l - r_{main} - \tilde{r}_{sub}) = r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}$ of the k separating slices can be assigned to E.

In addition, we assign to E the slices between lines representing the same main-only category in E and the slices between lines representing the same subcategory in E. This leads to

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}}{n} + \sum_{j \in OJ^*} \Big(\sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{k_j - 1}{n} \Big). \tag{5.9}$$

Example 5.1.6. Consider the data set described in Example 5.1.5. Suppose that we are interested in the event $Y_9 \in \{\text{LB,DB,LG}\}$. We have $k = 4$, $r_{main} = 0$, $r_{sub} = 2$, $r = 2$ and $l = 0$. For main categories with subcategories, the values of k_j, r_j and l_j are shown in Table 5.6. The values in Table 5.6 show that the condition

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

Table 5.6: Values of k_j, r_j and l_j for Example 5.1.6

$k_j - r_j + 1 \leq r_j + l_j$ is satisfied by main category B but not by main category G

and we have $\tilde{r}_{sub} = 1$. We also have $r^0_{sub} = 0$. This example illustrates the situation where $(k - r) + r^0_{sub} > l + r_{main} + \tilde{r}_{sub}$ and $k_j - r_j - 1 \leq r_j + l_j$ for all j. We can configure the probability wheel such that within each segment, every separating slice between subcategories is assigned to a subcategory in E. However, we cannot assign to E all separating slices between main categories. One configuration of the

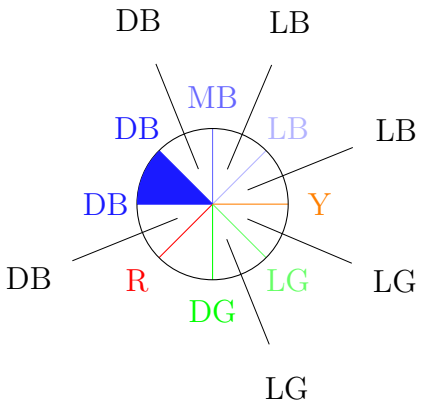

Figure 5.6: Probability wheel for Example 5.1.6

wheel corresponding to the NPI upper probability is shown in Figure 5.6, where R and Y are separated by B, and G and Y are separated by LG. However, there is no available main category or subcategory in E to which we can assign the slice separating R and G. This leads to a NPI upper probability of $\frac{7}{8}$ for the event E. This upper probability can be verified using (5.9). We see that the set OJ is empty, the set OJ^* contains B and G, the set OI_1 contains LB and DB and the set OI_2 contains LG. Also, $r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub} = 3$, therefore (5.9) gives

$$\overline{P}(E) = \frac{3}{8} + \frac{1-1}{8} + \frac{2-1}{8} + \frac{2}{8} + \frac{1-1}{8} + \frac{1}{8} = \frac{7}{8}.$$

◇

Case 3: $(k - r) + r^0_{sub} \leq l + r_{main} + \tilde{r}_{sub}$, $k_j - r_j - 1 > r_j + l_j$ for at least one j

In this situation, the number of main categories in E that can be used to separate observed main categories not in E is greater than or equal to the number of observed main categories with neither end of their segment in E. This means that

all k separating slices between main categories can be assigned to E.

However, within certain individual segments, we cannot separate all observed subcategories not in E. There will be $k_j - r_j - 1 - (r_j + l_j) = k_j - 2r_j - l_j - 1$ slices remaining. We can therefore only assign $k_j - 1 - (k_j - 2r_j - l_j - 1) = 2r_j + l_j$ separating slices to E.

We must also assign to E the slices which lie between lines representing the same main-only category in E and the slices which lie between lines representing the same subcategory in E. This leads to

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{k}{n} + \sum_{j \in OJ^*} \Big(\sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{\min\{2r_j + l_j, k_j - 1\}}{n} \Big). \quad (5.10)$$

Example 5.1.7. Consider the data set described in Example 5.1.5. Suppose that we are interested in the event $Y_9 \in \{\text{LB,DB,P,O}\}$. We have $k = 4$, $r_{main} = 0$, $r_{sub} = 1$, $r = 1$ and $l = 2$. The values of k_j, r_j and l_j are shown in Table 5.7. The values in

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

Table 5.7: Values of k_j, r_j and l_j for Example 5.1.7

Table 5.7 show that the condition $k_j - r_j - 1 \leq r_j + l_j$ is satisfied by main category B but not by main category G. Furthermore, the condition $k_j - r_j + 1 \leq r_j + l_j$ is only satisfied by B, so we have $\tilde{r}_{sub} = 1$. We also have $r^0_{sub} = 0$. This is an example of the situation where $(k - r) + r^0_{sub} \leq l + r_{main} + \tilde{r}_{sub}$ and $k_j - r_j - 1 > r_j + l_j$ for one or more j. We can configure the probability wheel such that all k separating slices between two lines representing different main categories can be assigned to E. However, within the G segment we cannot separate all subcategories not in E. A configuration of the wheel corresponding to the NPI upper probability is shown in Figure 5.7, where R and Y are separated by B, G and Y are separated by O, and

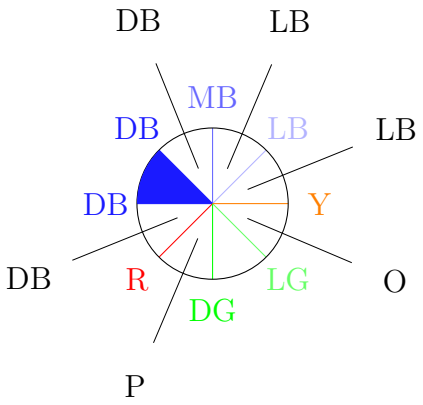

Figure 5.7: Probability wheel for Example 5.1.7

G and R are separated by P. However, there is no available subcategory in E to which we can assign the slice separating DG and LG. This leads to a NPI upper probability of $\frac{7}{8}$ for the event E. This upper probability can be verified using (5.10). We see that the set OJ is empty, the set OJ^* contains B and the set OI_1 contains LB and DB. Also, $\sum_{j \in OJ^*} \min\{2r_j + l_j, k_j - 1\} = 2$, therefore (5.10) gives

$$\overline{P}(E) = \frac{4}{8} + \frac{1-1}{8} + \frac{2-1}{8} + \frac{2}{8} = \frac{7}{8}.$$

◇

Case 4: $(k-r) + r^0_{sub} > l + r_{main} + \tilde{r}_{sub}$, $k_j - r_j - 1 > r_j + l_j$ for at least one j

In this situation, there are more observed main categories with neither end of their segment in E than there are main categories in E which are either main-only or have both ends of their segment in E. This means that, as in Case 2, only $r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}$ of the k separating slices between main categories can be assigned to E.

Also, within some particular main categories there are more observed subcategories not in E than there are subcategories in E, so as in Case 3 only $2r_j + l_j$ of the separating slices between subcategories can be assigned to E. Again, we must assign to E any slices between lines representing the same main-only category in E

and any slices between lines representing the same subcategory in E. This leads to

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}}{n} + \sum_{j \in OJ^*} \Big(\sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{\min\{2r_j + l_j, k_j - 1\}}{n} \Big). \tag{5.11}$$

Example 5.1.8. Consider the data set described in Example 5.1.5. Suppose that we are interested in the event $Y_9 \in \{\text{LB,DB,P}\}$. We have $k = 4$, $r_{main} = 0$, $r_{sub} = 1$, $r = 1$ and $l = 1$. For main categories described at subcategory level, the values of k_j, r_j and l_j are shown in Table 5.8. The values in Table 5.8 show that the condition

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

Table 5.8: Values of k_j, r_j and l_j for Example 5.1.8

$k_j - r_j - 1 \leq r_j + l_j$ is satisfied by main category B but not by main category G. We also see that the condition $k_j - r_j + 1 \leq r_j + l_j$ is only satisfied by B, so $\tilde{r}_{sub} = 1$. We also have $r^0_{sub} = 0$. We have the situation where $(k - r) + r^0_{sub} > l + r_{main} + \tilde{r}_{sub}$ and $k_j - r_j - 1 > r_j + l_j$ for one or more j. There is no configuration of the probability wheel such that all categories not in E are separated by categories in E. Also, within the G segment we cannot separate all subcategories not in E. One

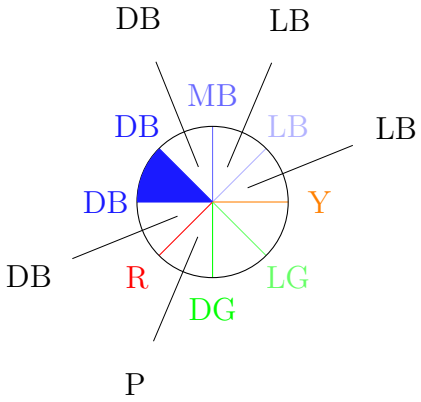

Figure 5.8: Probability wheel for Example 5.1.8

configuration of the wheel corresponding to the NPI upper probability is shown in Figure 5.8. Figure 5.8 shows a configuration where R and Y are separated by B, and G and R are separated by P. However, we cannot separate G and Y by a category in E. We also do not have an available subcategory in E to which we can assign the slice separating DG and LG. This leads to a NPI upper probability of $\frac{6}{8}$ for the event E. This upper probability can be verified using (5.11). We see that the set OJ is empty, the set OJ^* contains B and the set OI_1 contains LB and DB. Also, $r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub} = 3$ and $\sum_{j \in OJ^*} \min\{2r_j + l_j, k_j - 1\} = 2$, therefore (5.11) gives

$$\overline{P}(E) = \frac{3}{8} + \frac{1-1}{8} + \frac{2-1}{8} + \frac{2}{8} = \frac{6}{8}.$$

$\diamond$

The NPI upper probability formulae for the four cases in this subsection can be combined to give the following general expression:

$$\begin{aligned}\overline{P}(E) = &\sum_{j \in OJ} \frac{n_j - 1}{n} + \frac{\min\{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}, k\}}{n} \\ &+ \sum_{j \in OJ^*} \Big(\sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{\min\{2r_j + l_j, k_j - 1\}}{n} \Big).\end{aligned} \tag{5.12}$$

5.2 Properties of the model

In this section, we prove some general properties of the results presented in Section 5.1, where a general expression for the event of interest E was given (5.1) and the NPI lower probability (5.7) and upper probability (5.12) for this event were derived. We discuss four properties of these inferences.

5.2.1 Conjugacy

A fundamental property for lower and upper probabilities is the conjugacy property. Conjugacy means that for any event of interest E, the following expression is always true:

$$\overline{P}(E) = 1 - \underline{P}(E^c).$$

E^c represents the complementary event to E, i.e. an event containing all possible categories and subcategories except for those contained in E. We illustrate conjugacy in Example 5.2.1 below.

Example 5.2.1. Consider a multinomial data set with possible main categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). These are labelled 1 to 6 respectively. Observations in B are classified further as light blue (LB), medium blue (MB), dark blue (DB) or other blue (OB) and observations in G are classified further as light green (LG), dark green (DG) or other green (OG). The data set consists of eight observations, namely 1 LB, 1 MB, 2 DB, 1 LG, 1 DG, 1 R and 1 Y.

Suppose that we are interested in the event $Y_9 \in \{\text{MB,OB,DG,OG,R,Y,O}\}$. We label this event E. To find the upper probability for E we use (5.12). For this event, $K = 6$, $k = 4$, $r = 4$ and $l = 1$. For main categories with subcategories, the values of K_j, k_j, r_j, l_j, $\overline{r_j}$ and $\overline{l_j}$ are shown in Table 5.9. We have $r_{main} = 2$, $r_{sub} = 2$ and $\tilde{r}_{sub} = 1$. Substituting these values into (5.12) gives $\overline{P}(E) = \frac{7}{8}$.

	j	K_j	k_j	r_j	l_j	$\overline{r_j}$	$\overline{l_j}$
B	1	4	3	1	1	2	0
G	2	3	2	1	1	1	0

Table 5.9: Values for E for Example 5.2.1

	j	K_j	r_j	l_j	$\overline{r_j}$	$\overline{l_j}$
B	1	4	2	0	1	1
G	2	3	1	0	1	1

Table 5.10: Values for E^c for Example 5.2.1

We now consider the complementary event to E, which is $Y_9 \in \{\text{LB,DB,LG,P}\}$. We label this event E^c. To find the lower probability for E^c, we use (5.7). For E^c, $K = 6$, $r = 2$ and $l = 1$. For main categories described at subcategory level, the values of K_j, r_j, l_j, $\overline{r_j}$ and $\overline{l_j}$ are shown in Table 5.10. Substituting these values

into (5.7) gives $\underline{P}(E^c) = \frac{1}{8}$ and hence

$$\underline{P}(E^c) = \frac{1}{8} = 1 - \frac{7}{8} = 1 - \overline{P}(E).$$

This shows that the general formulae for NPI lower and upper probability are conjugated with regard to the specific event considered here. $\diamond$

We now show that the conjugacy property is satisfied for all events. We do this by considering the general expressions for E and E^c. We already have a general formula for E (5.1) and we can express the complementary event E^c as

$$Y_{n+1} \in \bigcup_{j \in F} c_j \cup \bigcup_{j \in J^*} \bigcup_{i_j \in \overline{I_j}} s_{ji_j} \tag{5.13}$$

where $F \subseteq \{1, ..., K\}$. Note that F, J and J^* form a partition of $\{1, ..., K\}$ and are pairwise disjoint and $\overline{I_j} = \{1, ..., K_j\} \backslash I_j$ for $j = 1, ..., K$ as before.

We let $OF = F \cap \{1, ..., k\}$ and define $\mid OF \mid = f_{main}$. OF is the index-set of observed main-only categories in E^c. Also, let $UF = F \cap \{k+1, ..., K\}$ and define $\mid UF \mid = q_{main}$. UF is the index-set of unobserved main-only categories in E^c. Table 5.11 shows how the notation used for the two events E and E^c compares. We take (5.7) and substitute E^c in place of E. This shows that the NPI lower

	E	E^c
Observed main categories	$r_{main} + r_{sub} = r$	$f_{main} + r_{sub}$
Unobserved main categories	$l_{main} + l_{sub} = l$	$q_{main} + l_{sub}$
Observed subcategories in event	r_j	$\overline{r_j}$
Unobserved subcategories in event	l_j	$\overline{l_j}$
Observed subcategories not in event	$\overline{r_j}$	r_j
Unobserved subcategories not in event	$\overline{l_j}$	l_j

Table 5.11: Notation for E and E^c

probability for E^c is

$$\begin{aligned}\underline{P}(E^c) = &\sum_{j\in OF}\frac{n_j-1}{n} + \sum_{j\in OJ^*}\sum_{i_j\in\overline{OI_j}}\frac{n_{j,i_j}-1}{n}\\ &+\frac{1}{n}(2f_{main}+2r_{sub}+q_{main}+l_{sub}-K-\overline{S_M})^+\\ &+\frac{1}{n}\sum_{j\in OJ^*}(2\overline{r_j}+\overline{l_j}-K_j-1)^+\end{aligned} \tag{5.14}$$

where $\overline{S_M} = \sum_{j\in OJ^*}\min\{(r_j+l_j-\overline{r_j}+1)^+,2\} + \sum_{j\in UJ^*}\min\{l_j,1\}$. To investigate the conjugacy property, we compare (5.14) to the NPI upper probability (5.12) for E .

Theorem 5.2.1. $\underline{P}(E^c) = 1-\overline{P}(E)$.

Proof. To satisfy conjugacy, $\underline{P}(E^c)$ and $\overline{P}(E)$ must always sum to 1. This means for the probability wheel that the total number of slices of the wheel assigned to E according to $\overline{P}(E)$ or to E^c according to $\underline{P}(E^c)$ must sum to n.

Within each individual segment, there are n_j-1 slices in total including $n_{j,i_j}-1$ slices which must belong to each subcategory s_{j,i_j}. We need to prove that the total number of separating slices assigned to E or E^c in an individual segment always sums to k_j-1. The formula for $\underline{P}(E^c)$ assigns $(2\overline{r_j}+\overline{l_j}-K_j-1)^+$ of these separating slices to E^c and the formula for $\overline{P}(E)$ assigns $\min\{2r_j+l_j, k_j-1\}$ of these separating slices to E. So we need to show two things: first, that $2\overline{r_j}+\overline{l_j}-K_j-1$ and $2r_j+l_j$ sum to k_j-1 and secondly, that if $k_j-1<2r_j+l_j$, then $2\overline{r_j}+\overline{l_j}-K_j-1<0$.

Since $r_j+\overline{r_j}=k_j$, and $r_j+\overline{r_j}+l_j+\overline{l_j}=K_j$, we have

$$2\overline{r_j}+\overline{l_j}-K_j-1+2r_j+l_j = k_j+K_j-K_j-1 = k_j-1.$$

Also,

$$k_j-1<2r_j+l_j \Leftrightarrow k_j-1-2r_j-l_j<0 \Leftrightarrow \overline{r_j}-r_j-l_j-1<0 \Leftrightarrow 2\overline{r_j}+\overline{l_j}-K_j-1<0.$$

So this shows that the total number of slices in a segment will always sum to n_j-1 as required. This is also clearly true for main-only categories.

Looking at the wheel as a whole, we need to show that the number of separating slices between main categories that are assigned to E or E^c always sums to k. The formula for $\overline{P}(E)$ assigns $\min\{r+l+r_{main}+\tilde{r}_{sub}-r^0_{sub}, k\}$ of these separating slices to E and the formula for $\underline{P}(E^c)$ assigns $(2f_{main}+2r_{sub}+q_{main}+l_{sub}-K-\overline{S_M})^+$ of these separating slices to E^c. So we need to show two things: first, that $r+l+r_{main}+\tilde{r}_{sub}-r^0_{sub}$ and $2f_{main}+2r_{sub}+q_{main}+l_{sub}-K-\overline{S_M}$ sum to k and secondly, that if $r+l+r_{main}+\tilde{r}_{sub}-r^0_{sub} > k$ then $2f_{main}+2r_{sub}+q_{main}+l_{sub}-K-\overline{S_M} < 0$.

By considering the definitions of $\tilde{r}_{sub}$ and r^1_{sub}, we find that

$$\overline{S_M} = 2\tilde{r}_{sub} + r^1_{sub} + l_{sub}.$$

Using this and also using the relations $r_{main}+f_{main}+r_{sub}+l_{main}+q_{main}+l_{sub}=K$ and $r_{main}+f_{main}+r_{sub}=k$, we have

$$r+l+r_{main}+\tilde{r}_{sub}-r^0_{sub}+2f_{main}+2r_{sub}+q_{main}+l_{sub}-K-\overline{S_M} =$$
$$r+l+r_{main}-r_{sub}+2f_{main}+2r_{sub}+q_{main}-K = r_{main}+r_{sub}+f_{main} = k.$$

Also,

$$r+l+r_{main}+\tilde{r}_{sub}-r^0_{sub} > k \Leftrightarrow k-r-l-r_{main}-\tilde{r}_{sub}+r^0_{sub} < 0$$
$$\Leftrightarrow 2f_{main}+2r_{sub}+q_{main}+l_{sub}-K-\overline{S_M} < 0.$$

This proves that the number of separating slices between main categories which are assigned to E or E^c always sums to k. This completes the proof that the total number of slices assigned to E and E^c is always equal to n, or in other words that $\overline{P}(E)$ and $\underline{P}(E^c)$ sum to 1. Therefore the general formulae (5.7) and (5.12) for the NPI lower and upper probability are conjugated, as required. □

5.2.2 Relative frequencies

A desirable property for the Sub-MNPI model is that the interval between the lower and upper probabilities contains the relative frequency of observations in the event of interest E. We label the relative frequency as RF, where

$$RF = \sum_{j\in OJ}\frac{n_j}{n} + \sum_{j\in OJ^*}\sum_{i_j\in OI_j}\frac{n_{j,i_j}}{n}. \tag{5.15}$$

We want to prove that

$$\underline{P}(E) \leq RF \leq \overline{P}(E). \tag{5.16}$$

A basic form of predictive inference is the method whereby the predictive probability of a category is simply set to be to the relative frequency of that category in the data. The above property (5.16) is desirable because we wish to show that the Sub-MNPI model is a generalisation of this intuitive method and furthermore that the lower and upper probabilities given by the model are not in conflict with the empirical probabilities. This cannot be said of precise probability methods such as Bayesian inferences, which typically assign a positive probability to a category before it has been observed even once. Example 5.2.2 illustrates (5.16).

Example 5.2.2. Consider a multinomial data set with possible main categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). Observations in B are classified further as light blue (LB), medium blue (MB), dark blue (DB) or other blue (OB) and observations in G are classified further as light green (LG), dark green (DG) or other green (OG). The data set consists of eight observations, namely 1 LB, 1 MB, 2 DB, 1 LG, 1 DG, 1 R and 1 Y.

S	$\underline{P}$	RF	$\overline{P}$
P	0	0	$\frac{1}{8}$
P,LB	0	$\frac{1}{8}$	$\frac{3}{8}$
P,LB,LG	0	$\frac{2}{8}$	$\frac{5}{8}$
P,LB,LG,DB	$\frac{1}{8}$	$\frac{4}{8}$	1
P,LB,LG,DB,R	$\frac{1}{8}$	$\frac{5}{8}$	1
P,LB,LG,DB,R,Y	$\frac{1}{8}$	$\frac{6}{8}$	1
P,LB,LG,DB,R,Y,MB	$\frac{1}{8}$	$\frac{7}{8}$	1
P,LB,LG,DB,R,Y,MB,DG	$\frac{1}{8}$	1	1

Table 5.12: Lower and upper probabilities and relative frequencies for Example 5.2.2

Several general events of the form $Y_9 \in S$ are considered here, where S is a subset of the K possible categories. These are shown in Table 5.12, together with the corresponding relative frequencies and NPI lower and upper probabilities, which are

found using (5.15), (5.7) and (5.12) respectively. We see that (5.16) is true for all events considered in this example. ◇

We now prove (5.16) for the general case. We prove this property in two stages, considering each inequality separately.

Theorem 5.2.2. $\underline{P}(E) \leq RF$.

Proof. We can rearrange the formula for $\underline{P}(E)$ (5.7) as

$$\begin{aligned}\underline{P}(E) = &\sum_{j\in OJ} \frac{n_j}{n} - \frac{r_{main}}{n} + \sum_{j\in OJ^*} \Big(\sum_{i_j \in OI_j} \frac{n_{j,i_j}}{n} - \frac{r_j}{n} \Big) \\ &+ \frac{1}{n}(2r + l - K - S_M)^+ \\ &+ \frac{1}{n} \sum_{j\in OJ^*} (2r_j + l_j - K_j - 1)^+.\end{aligned} \tag{5.17}$$

In the event that the terms $(2r+l-K-S_M)^+$ and $(2r_j+l_j-K_j-1)^+$ are both equal to zero, the proof is trivial. Supposing that they are both non-zero, the inequality that must be proven is equivalent to

$$-r_{main} + 2r + l - K + \sum_{j\in OJ^*} (r_j + l_j - K_j - 1) - S_M \leq 0.$$

Since $\sum_{j\in OJ^*} 1 = r_{sub}$, the inequality becomes

$$(r + l - K) + \sum_{j\in OJ^*} (r_j + l_j - K_j) - S_M \leq 0.$$

By definition, $(r + l - K)$ and $(r_j + l_j - K_j)$ must be less than or equal to zero, and the third term always subtracts a number greater than or equal to zero, therefore we see that the LHS is always non-positive. This proves the required inequality.

If $(2r_j + l_j - K_j - 1)^+$ is equal to zero but $(2r + l - K - S_M)^+$ is not, the inequality that must be proven is equivalent to

$$-r_{main} + 2r + l - K - \sum_{j\in OJ^*} r_j - S_M \leq 0.$$

The term $\sum_{j\in OJ^*} r_j + S_M$ is always at least as large as $\sum_{j\in OJ^*} 1 = r_{sub}$ and this leads to the inequality

$$r + l - K \leq 0$$

which is true by definition.

If $(2r + l - K - S_M)^+$ is equal to zero but $(2r_j + l_j - K_j - 1)^+$ is not, the inequality that must be proven is equivalent to

$$-r_{main} + \sum_{j \in OJ^*} (-r_j + 2r_j + l_j - K_j - 1) \leq 0$$

which can be rewritten as

$$-r_{main} + \sum_{j \in OJ^*} (r_j + l_j - K_j - 1) \leq 0.$$

By definition, $(r_j + l_j - K_j)$ is less than or equal to zero, so the LHS is always non-positive. This proves the required inequality. So Theorem 5.2.2 holds in all situations. □

Theorem 5.2.3. *$RF \leq \overline{P}(E)$.*

Proof. The formula for $\overline{P}(E)$ can be rewritten as

$$\begin{aligned} \overline{P}(E) = &\sum_{j \in OJ} \frac{n_j}{n} - \frac{r_{main}}{n} + \frac{\min\{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}, k\}}{n} \\ &+ \sum_{j \in OJ^*} \Big(\sum_{i_j \in OI_j} \frac{n_{j,i_j}}{n} - \frac{r_j}{n} + \frac{\min\{2r_j + l_j, k_j - 1\}}{n} \Big) \end{aligned} \tag{5.18}$$

and so the inequality that must be proven is equivalent to

$$0 \leq -r_{main} + \min\{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}, k\} + \sum_{j \in OJ^*} (-r_j + \min\{2r_j + l_j, k_j - 1\}).$$

Since $k_j = r_j + \overline{r_j}$, this inequality becomes

$$0 \leq \min\{r + l + \tilde{r}_{sub} - r^0_{sub}, k - r_{main}\} + \sum_{j \in OJ^*} \min\{r_j + l_j, \overline{r_j} - 1\}.$$

The RHS is always positive, proving Theorem 5.2.3. □

Combining Theorems 5.2.2 and 5.2.3 shows that the model contains the relative frequencies, so (5.16) is satisfied.

5.2.3 Imprecision vanishes as $n \to \infty$

A third property of the Sub-MNPI model is that as the number of observations in the data set becomes infinitely large, the imprecision vanishes and the interval probability $P(E)$ shrinks to a point value equal to the relative frequency, which is given by (5.15).

Theorem 5.2.4. $\lim_{n\to\infty}[\overline{P}(E) - \underline{P}(E)] = 0.$

Proof. Using (5.7) and (5.12), $\overline{P}(E) - \underline{P}(E)$ can be written as

$$\begin{aligned}&\frac{1}{n}(\min\{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}, k\}) + \frac{1}{n}\sum_{j\in OJ^*}(\min\{2r_j + l_j, k_j - 1\} - r_j)\\ &- \frac{1}{n}(2r + l - K - S_M)^+ - \frac{1}{n}\sum_{j\in OJ^*}((2r_j + l_j - K_j - 1)^+ - r_j).\end{aligned} \tag{5.19}$$

We consider what happens to the terms in this expression as $n \to \infty$. The value of k increases as more data are observed because the number of main categories that have not been observed diminishes and we find that $k \to k^\infty$ for some $k^\infty \leq K$. The same applies to the total number of subcategories and we find that $k_j \to k_j^\infty$ for some $k_j^\infty \leq K_j$. Similarly, $r, l, r_j, l_j, \overline{r_j}$ and $\overline{l_j}$ are affected: as the numbers of observed main categories and subcategories increase, r, r_j and $\overline{r_j}$ increase whilst l, l_j and $\overline{l_j}$ decrease. However, all these terms remain finite, since they are limited by the total numbers of main categories and subcategories. The total number of main categories, K, and the total number of subcategories in each main category, K_j, are unchanged as n increases and remain finite. (5.19) tends to zero as $n \to \infty$, so $\lim_{n\to\infty}[\overline{P}(E) - \underline{P}(E)] = 0.$ □

5.2.4 F-probability

An important property for the Sub-MNPI model to satisfy is that the interval probabilities $[\underline{P}(E), \overline{P}(E)]$ are F-probabilities in the sense of Weichselberger [43].

Let Ω represent the sample space consisting of all possible main categories and subcategories and suppose that $\mathcal{E}$ is the power set of Ω. An event E is an element of $\mathcal{E}$. Weichselberger [43] defines the structure as the set of all classical

probabilities p that are in accordance with the interval limits, i.e. the set

$$\mathcal{M} = \{p | \underline{P}(E) \leq p(E) \leq \overline{P}(E), \forall E \in \mathcal{E}\}.$$

An interval probability $P(\cdot)$ is an F-probability if, for all $E \in \mathcal{E}$,

$$\inf_{p \in \mathcal{M}} p(E) = \underline{P}(E)$$

and

$$\sup_{p \in \mathcal{M}} p(E) = \overline{P}(E).$$

F-probability is a desirable property because it shows that none of the interval probabilities is too wide and that they could not be made any smaller given the data available to us. Also, F-probability is strongly linked to other concepts in probability theory. Conjugacy, proven in Subsection 5.2.1, is implicit in the F-probability property. Coherence is a direct consequence of F-probability, by Walley's lower envelope theorem [40], and this can be seen as a rationality requirement. From a subjective perspective, lower and upper probabilities can be used to determine betting behaviour, and coherence ensures rational behaviour and that no sure-loss gambles are accepted. It should be noted that the NPI lower and upper probabilities are not coherent in the sense of Walley upon updating, but they are at any single point in time. Updating, however, is not done in a Bayesian way through conditioning as Walley's coherence implies, but is dealt with by taking new observations into account as well as the previously observed data and making inferences based on the new total number of observations. We do not consider updating or conditioning here, but in [6] and [18] it is shown that NPI leads to strong consistency properties for these actions.

In order to investigate the F-probability property, we introduce some new notation to describe all the possible configurations of the probability wheel. As stated previously, there are K main categories in total and there are K_j subcategories within each category. We now imagine that the wheel is split into K segments and that each segment is split into K_j subsegments. We move clockwise around the wheel numbering the segments as $1, ..., K$ as shown in Figure 5.9. We

also number the subsegments within segment j as $1, ..., K_j$ as shown in Figure 5.10. The area of these segments and subsegments is thus far unspecified: we allocate a different main category or subcategory to each segment or subsegment in order to describe the configuration of the wheel, but a segment assigned to an unobserved category may have area zero.

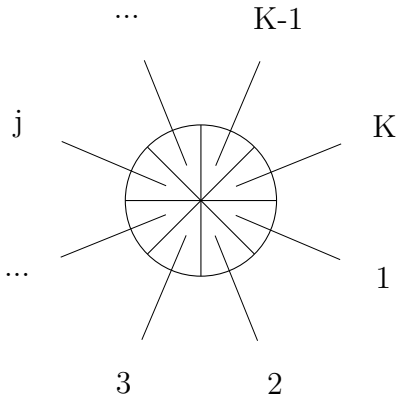

Figure 5.9: Diagram showing how segments are numbered

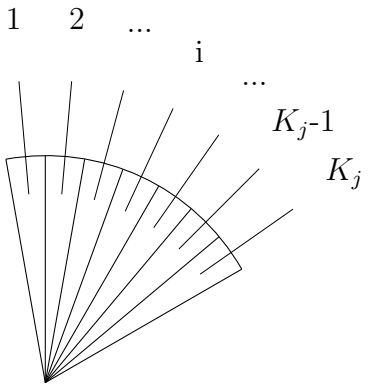

Figure 5.10: Diagram showing how subsegments are numbered

As seen in [17], we let Σ represent the set of all possible permutations σ of the wheel. Each σ can be described by a sequence

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

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

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

where $\sigma(i,j)$ is the index of the subcategory within main category j assigned to subsegment i.

It is also necessary to describe the position of the observed main categories and subcategories on the wheel for a given σ. Let the circular sequence $\sigma(i_1),...,\sigma(i_{k+1})$, with $\sigma(i_{k+1}) = \sigma(i_1)$, be the indices of the observed main categories as we move around the wheel and let the sequence $\sigma(i_1,j),...,\sigma(i_{k_j},j)$, $j \in J^*$, be the indices of the observed subcategories as we move through the segment representing main category j.

Using the above notation, Coolen and Augustin [17] described the separating slice between two observed main categories using the set $\{\sigma(j)|i_l \leq j \leq i_{l+1}\}$ for $l = 1,...,k$. This is the set of indices of all possible main categories to which, for the particular configuration σ, we could assign the separating slice between the main categories in positions i_l and i_{l+1} on the wheel.

However, since we are now considering the situation where main categories may be broken down into subcategories, we must also consider the specific subcategories to which the slice could be assigned. We describe the separating slice as follows:

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

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

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

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

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

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

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

if categories in positions i_l and i_{l+1} both have subcategories.

$J_{\sigma,l}$ is the index set of all main categories and subcategories to which the separating slice could be assigned. Let $c_{|J_{\sigma,l}|}$ be the set of all these main categories and subcategories.

We also describe the separating slice between two observed subcategories within the same main category using

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

This is the set of indices of all possible subcategories to which, for the particular configuration σ, we could assign the separating slice between the subcategories in positions i_l and i_{l+1} in the segment representing main category j. Let $s_{|B_{\sigma,j,l}|}$ be the set of these subcategories.

Example 5.2.3 is included to clarify the above notation.

Example 5.2.3. Consider a multinomial data set with possible main categories red (R), yellow (Y), blue (B), green (G), pink (P), purple (Pu), orange (O) and white (W). These categories are numbered 1 to 8 respectively. We can describe B in more detail: observations may be dark blue (DB), medium blue (MB) or light blue (LB). These subcategories are numbered 1 to 3 respectively. Similarly, G observations may be dark green (DG), medium green (MG) or light green (LG). These subcategories are numbered 1 to 3 respectively.

First we consider the positioning of these main categories and subcategories on the wheel. As explained above, there is a set of configurations Σ detailing the position of every possible main category and subcategory, both observed and unobserved. Suppose that we are looking at one particular configuration, σ. The positions of the main categories and subcategories in this configuration are as follows:

$\sigma(1) = 1,\ \sigma(2) = 5,\ \sigma(3) = 3,\ \sigma(4) = 6,\ \sigma(5) = 4,\ \sigma(6) = 7,\ \sigma(7) = 2,$

$\sigma(8) = 8$

$\sigma(1,3) = 1,3,\ \sigma(2,3) = 2,3,\ \sigma(3,3) = 3,3$

$\sigma(1,4) = 3,4,\ \sigma(2,4) = 2,4,\ \sigma(3,4) = 1,4$

For further clarification, σ is shown in Figure 5.11.

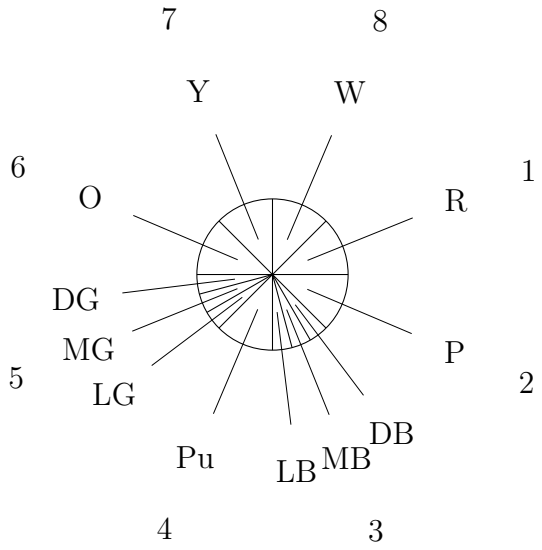

Figure 5.11: Configuration σ for Example 5.2.3

Suppose that we have observed 2 R, 1 Y, 2 DB, 1 LB, 1 MG and 1 DG. Given the configuration σ explained above, the probability wheel for these data is as shown in Figure 5.12. In order to decide how to assign each of the slices separating different main categories, we use the sets $J_{\sigma,l}$ explained above. The slice labelled S1, which separates the Y and R main categories, is an example of a slice separating two main-only categories. We therefore use the formula $J_{\sigma,l} = \{\sigma(j)|i_l \leq j \leq i_{l+1}\}$. Here, $J_{\sigma,l} = \{\sigma(j)|7 \leq j \leq 9\} = \{\sigma(7), \sigma(8), \sigma(9)\}$. So this slice could be assigned to R, W or Y.

The slice labelled S2, which separates the R and B main categories, is an example of a separating slice where the category in position i_l is main-only but the category in position i_{l+1} has subcategories. We therefore use the formula

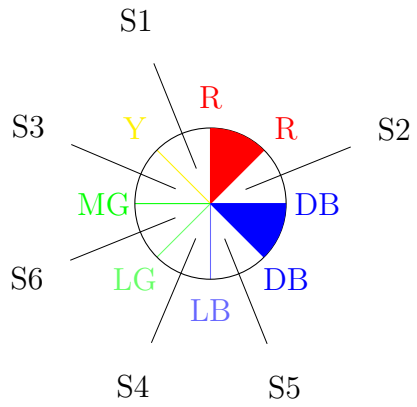

Figure 5.12: Probability wheel for Example 5.2.3

$J_{\sigma,l} = \{\sigma(j)|i_l \leq j < i_{l+1}\} \cup \bigcup_{x=1}^{i_1} \sigma(x, l+1)$. Here, $J_{\sigma,l} = \{\sigma(j)|1 \leq j < 3\} \cup \sigma(1,3)$ $= \{\sigma(1), \sigma(2), \sigma(1,3)\}$. So this slice could be assigned to R, P or DB.

The slice labelled S3, which separates the G and Y main categories, is an example of a separating slice where the category in position i_l has subcategories but the category in position i_{l+1} is main-only. We therefore use the formula $J_{\sigma,l} = \{\sigma(j)|i_l < j \leq i_{l+1}\} \cup \bigcup_{x=i_{k_l}}^{K_l} \sigma(x, l)$. Here, $J_{\sigma,l} = \{\sigma(j)|5 < j \leq 7\} \cup \bigcup_{x=2}^{3} \sigma(x, 5)$ $= \{\sigma(6), \sigma(7), \sigma(2,5), \sigma(3,5)\}$. So this slice could be assigned to O, Y, MG or DG.

The slice labelled S4, which separates the B and G main categories, is an example of a separating slice where the categories in positions i_l and i_{l+1} both have subcategories. We therefore use the formula $J_{\sigma,l} = \{\sigma(j)|i_l < j < i_{l+1}\} \cup \bigcup_{x=1}^{i_1} \sigma(x, l+1) \cup \bigcup_{x=i_{k_l}}^{K_l} \sigma(x, l)$. Here, $J_{\sigma,l} = \{\sigma(j)|3 < j < 5\} \cup \sigma(1,5) \cup \sigma(3,3) = \{\sigma(4), \sigma(1,5), \sigma(3,3)\}$. So this slice could be assigned to Pu, LB or LG.

In order to decide how to assign each of the slices separating different subcategories within a segment, we use the sets $B_{\sigma,j,l}$, also explained above.

The slice labelled S5, which separates the DB and LB subcategories, could be assigned to any subcategory in $B_{\sigma,j,l} = \{\sigma(b,3)|1 \leq b \leq 3\} =$

$\{\sigma(1,3), \sigma(2,3), \sigma(3,3)\}$, i.e. to LB, MB or DB.

The slice labelled S6, which separates the LG and MG subcategories, could be assigned to any subcategory in $B_{\sigma,j,l} = \{\sigma(b,4)|1 \leq b \leq 2\} = \{\sigma(1,4), \sigma(2,4)\}$, i.e. to LG or MG. $\diamond$

Now that the notation has been explained, we prove the F-probability property. The following theorem and proof is based on work by Coolen and Augustin [17] which proved the F-probability property for the original MNPI model. The proof below generalises this for the Sub-MNPI model. In this proof, basic probability assignments are used to describe our inferences in a formal way. For some event A, the basic probability assignment $m(A)$ is a measure interpreted as the proportion of the available information that supports the claim that A occurs without assuming any particular division among proper subsets of A. The theory of basic probability assignments is summarised in [5]. As seen in [17], for a given configuration σ, the Sub-MNPI model gives the following basic probability assignment to the event $Y_{n+1} \in c_j$:

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

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

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

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

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

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

Any other event is given the basic probability assignment of zero.

The proof of Theorem 5.2.5 also utilises the belief function and the plausibility function, which draw on the Dempster-Shafer theory [37]. (Note that we do not make use of any further concepts from Dempster-Shafer theory, particularly not

their updating rules.) Let X_E represent the index set of the event of interest E. This set contains some one-dimensional elements, corresponding to main-only categories, and some two-dimensional elements, corresponding to subcategories. We also define a second event of interest $Y_{n+1} \in D$, where D is some subset of main categories and subcategories corresponding to an index set X_D.

The belief function [37] of a set is defined as the sum of the basic probability assignments of all possible subsets of that set. For a given configuration σ, the belief function of E is

$$\underline{P}_\sigma(E) = \sum_{X_D \subseteq X_E} m_\sigma(Y_{n+1} \in D). \tag{5.20}$$

The plausibility function [37] of a set is defined as the sum of the basic probability assignments of all possible sets which intersect that set. For a given configuration σ, the plausibility function of E is

$$\overline{P}_\sigma(E) = \sum_{X_D \cap X_E \neq \emptyset} m_\sigma(Y_{n+1} \in D). \tag{5.21}$$

Theorem 5.2.5. *The interval consisting of the lower probability (5.7) and the upper probability (5.12) is an F-probability in the sense of Weichselberger [43].*

Proof. In order to prove this property, we determine the lower and upper probabilities for event E via the belief function (5.20) and the plausibility function (5.21). We first derive the belief function of E. With the basic probability assignments explained above, the only subsets of E which have a non-zero basic probability assignment are single main-only categories or subcategories in E, or sets $c_{|J_{\sigma,l}|}$ where $J_{\sigma,l} \subseteq X_E$, or sets $s_{|B_{\sigma,j,l}|}$ where $B_{\sigma,j,l} \subseteq I_j$. This leads to the following belief function:

$$\begin{aligned}\underline{P}_\sigma(E) = &\sum_{j \in J} m_\sigma(\{Y_{n+1} \in c_j\}) + \sum_{j \in J^*} \sum_{i_j \in I_j} m_\sigma(\{Y_{n+1} \in s_{j,i_j}\}) \\ &+ \sum_{J_{\sigma,l} \subseteq X_E} m_\sigma(\{Y_{n+1} \in c_{|J_{\sigma,l}|}\}) + \sum_{B_{\sigma,j,l} \subseteq I_j} m_\sigma(\{Y_{n+1} \in s_{|B_{\sigma,j,l}|}\}).\end{aligned} \tag{5.22}$$

We now derive the plausibility function of E. With the basic probability assignments explained above, the only subsets of E which have a non-zero basic probability

assignment are single main-only categories or subcategories in E, or sets $c_{|J_{\sigma,l}|}$ where $J_{\sigma,l} \cap X_E \neq \emptyset$, or sets $s_{|B_{\sigma,j,l}|}$ where $B_{\sigma,j,l} \cap I_j \neq \emptyset$. This leads to the following plausibility function:

$$\begin{aligned}\overline{P}_\sigma(E) = &\sum_{j\in J} m_\sigma(\{Y_{n+1} \in c_j\}) + \sum_{j\in J^*}\sum_{i_j\in I_j} m_\sigma(\{Y_{n+1} \in s_{j,i_j}\}) \\ &+ \sum_{J_{\sigma,l}\cap X_E\neq\emptyset} m_\sigma(\{Y_{n+1} \in c_{|J_{\sigma,l}|}\}) + \sum_{B_{\sigma,j,l}\cap I_j\neq\emptyset} m_\sigma(\{Y_{n+1} \in s_{|B_{\sigma,j,l}|}\}).\end{aligned} \tag{5.23}$$

We have a set of belief functions and a set of plausibility functions corresponding to the set Σ of possible configurations of the probability wheel. In Section 5.1, we derived the lower and upper probability formulae of the Sub-MNPI model by considering all possible configurations $\sigma \in \Sigma$, resulting in

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

and

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

In other words, we took the lower and upper envelopes over all possible configurations.

According to Theorem 3.2 of [5], taking the lower and upper envelopes over all possible configurations leads to F-probability. Therefore the interval probability is an F-probability as required. □

5.3 Unknown numbers of (sub)categories

In this section we consider the situation where the quantities K and K_j, $j = 1, ..., K$, are unknown. It is important to note that in addition to these quantities being unknown, they are not assumed to have a finite limit. As in Section 5.1, we have observed k main categories and within category c_j we have observed k_j subcategories. In order to describe the general events of interest in this situation, we introduce some new notation. As in Subsection 2.2.2, let c_{j_s}, $s = 1, ..., r'$, be the observed main-only categories in the event of interest, let UN be the set of

Unobserved New main categories, which refers to any not yet observed category, and let DN_j, $j = 1, ..., l$, be the set of Defined New main categories, which is a subset of UN and which represents categories we wish to specify in the event of interest but which have not yet been observed.

In addition, let c_{j_s}, $s = r' + 1, ..., r$, be the observed main categories in the event of interest which are described at subcategory level and let $s_{j_s,i_{j_s}}$, $s = r' + 1, ..., r$, $i_{j_s} = 1, ..., r_s$, be the observed subcategories in the event of interest. Let $\tilde{DN}_{j_s,i_{j_s}}$, $i_{j_s} = 1, ..., d_s$, be the set of Defined New subcategories within the observed main categories c_{j_s} and let DN_{j,i_j}, $j = 1, ..., l$, $i_j = 1, ..., l_j$, be the set of Defined New subcategories within the Defined New main categories. Let $\tilde{UN}_{j_s}$, $s = 1, ..., r$ be the set of all Unobserved New subcategories within the observed main categories c_{j_s} and let UN_j, $j = 1, ..., l$ be the set of all Unobserved New subcategories within the Defined New main categories. A given event can be expressed as a union involving some or all of the above terms. Let $A, B \subseteq \{1, ..., k\}$ such that $A \cap B = \emptyset$. Any event of interest can be expressed using one of the two formulae shown below. The first general event is

$$\begin{aligned} Y_{n+1} \in & \bigcup_{s=1}^{r'} c_{j_s} \cup \bigcup_{s=r'+1}^{r} \Big(\bigcup_{i_{j_s}=1}^{r_s} s_{j_s,i_{j_s}} \Big) \\ & \cup \bigcup_{s\in A} \Big(\tilde{UN}_{j_s} \setminus \bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \cup \bigcup_{s\in B} \Big(\bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\ & \cup \bigcup_{j=1}^{l'} \Big(UN_j \setminus \bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \cup \bigcup_{j=l'+1}^{l} \Big(\bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big). \end{aligned} \tag{5.24}$$

The second general event is

$$\begin{aligned} Y_{n+1} \in & \bigcup_{s=1}^{r'} c_{j_s} \cup \bigcup_{s=r'+1}^{r} \Big(\bigcup_{i_{j_s}=1}^{r_s} s_{j_s,i_{j_s}} \Big) \\ & \cup \bigcup_{s\in A} \Big(\tilde{UN}_{j_s} \setminus \bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \cup \bigcup_{s\in B} \Big(\bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\ & \cup UN \setminus \Big\{ \bigcup_{j=1}^{l'} \Big(UN_j \setminus \bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \cup \bigcup_{j=l'+1}^{l} \Big(\bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \Big\}. \end{aligned} \tag{5.25}$$

We denote these by E_1 (5.24) and E_2 (5.25). The notation used in these expressions

allows us to include only defined subcategories of some main categories, but all undefined subcategories of other main categories. E_1 encompasses all events containing only defined main categories, whilst E_2 encompasses all events containing the set of undefined main categories. We can therefore describe any event using either (5.24) or (5.25). We now derive formulae for the NPI lower and upper probabilities for each of these general events.

5.3.1 Lower probability

We consider the NPI lower probabilities for events E_1 (5.24) and E_2 (5.25). These are found by minimising the number of slices of the probability wheel that must be assigned to the event of interest. The slices of the wheel which must always be assigned to an event of interest are those between two lines representing the same observed main-only category or subcategory in that event. So for both E_1 and E_2, the term

$$\bigcup_{s=1}^{r'} c_{j_s} \cup \bigcup_{s=r'+1}^{r} \left(\bigcup_{i_{j_s}=1}^{r_s} s_{j_s,i_{j_s}} \right),$$

always contributes to the NPI lower probability. This term includes specified observed main-only categories and specified observed subcategories.

Let n_{j_s} represent the number of times main category c_{j_s} has been observed. We also define $n_{j_s,i_{j_s}}$ to be the number of times we have observed subcategory $s_{j_s,i_{j_s}}$.

The lower probability formulae for E_1 and E_2 therefore include the term

$$\sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s=r'+1}^{r} \left(\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n} \right) \tag{5.26}$$

We also have to consider the separating slices of the wheel, i.e. the slices which separate different observed main categories and the slices which separate different observed subcategories within one main category segment. We want to assign as few of these slices as possible to the event of interest.

First we consider event E_1. This event includes only a finite number of unobserved main categories, and since we do not assume an upper bound on the total number of unobserved categories, we can assign any slices separating observed main categories either to unobserved main categories not in E_1 or to observed main categories not in E_1. For slices separating subcategories, the situation is more complicated. For each main category described at subcategory level, there are r_s observed subcategories in E_1, so to fully separate these we need at least $r_s - 1$ subcategories not in E_1. If $s \notin A$ (where $A \subseteq \{1, ..., k\}$), the main category in question only has a finite number of unobserved subcategories in E_1, so we can assign slices separating observed subcategories either to unobserved subcategories not in E_1 or to observed subcategories not in E_1. However, if $s \in A$, then there are only d_s unobserved subcategories which are not in E_1. We may have to assign some of the separating slices to subcategories in E_1 if there is an insufficient number of subcategories not in E_1. The number of separating slices we have to assign to E_1 is

$$N_s = [(r_s - 1) - d_s - (k_{j_s} - r_s)]^+. \tag{5.27}$$

So, given all of the above reasoning, the NPI lower probability for the event E_1 is

$$\begin{aligned}\underline{P}(E_1) = &\sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n}) \} \\ &+ \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n}) + \frac{N_s}{n} \}.\end{aligned} \tag{5.28}$$

Example 5.3.1. Consider a multinomial data set where the set of possible main categories consists of an unknown number of different colours. We have observed the following main categories: red (R), blue (B), green (G) and pink (P). At subcategory level, we have observed dark blue (DB), medium blue (MB), light blue (LB), dark green (DG), medium green (MG), light green (LG), medium pink (MP) and dark pink (DP). In addition we define two new main categories: orange (O), with defined subcategories light orange (LO) and medium orange (MO), and purple (Pu) with defined subcategory dark purple (DPu). We also define the new subcategory light pink (LP). We let UN_B represent all unobserved new subcategories within the main category B, including the defined new subcategory royal blue (RB), and let UN_{Pu}

represent the equivalent for the main category Pu. The data set consists of twenty observations including 3 R, 3 DB, 1 MB, 2 LB, 3 DG, 2 MG, 2 LG, 2 MP and 2 DP.

	B	G	P
n_{j_s}	6	7	4

Table 5.13: Values of n_{j_s} for Example 5.3.1

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

Table 5.14: Values of $n_{j_s,i_{j_s}}$ for Example 5.3.1

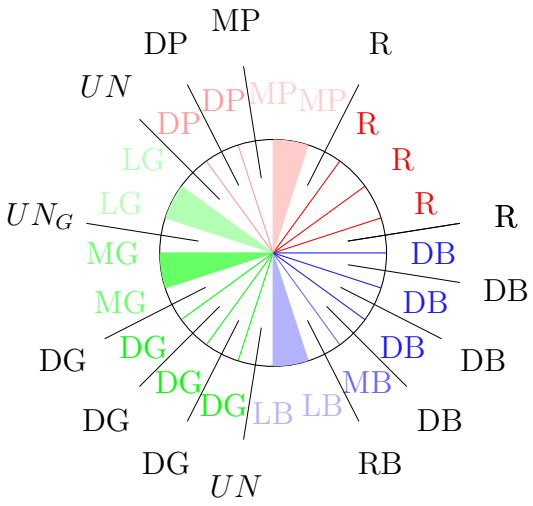

Figure 5.13: Probability wheel for Example 5.3.1

Suppose that we are interested in the event $Y_{21} \in \{(\text{LB} \cup \text{MB}) \cup (\text{LG} \cup \text{MG}) \cup (\text{MP}) \cup (UN_B \backslash \text{RB}) \cup (\text{LP}) \cup (UN_{Pu} \backslash \text{DPu}) \cup (\text{LO} \cup \text{MO})\}$. We label this event E. Let $s = 1$ correspond to B, $s = 2$ to G and $s = 3$ to P. Comparing E to the general formulae, (5.24) and (5.25), we see that this is an event of type E_1, so (5.28) is used to compute the NPI lower probability.

In this example, $r = 3$. The main categories for which $s \notin A$ are G and P and the only main category for which $s \in A$ is B. We have

$N_1 = [(r_1 - 1) - d_1 - (k_{j_1} - r_1)]^+ = [(2 - 1) - 1 - (3 - 2)]^+ = 0$. The values of n_{j_s} and $n_{j_s,i_{j_s}}$ are shown in Tables 5.13 and 5.14.

Putting these values into (5.28) shows that the NPI lower probability for the event E is

$$(\frac{2-1}{20}) + (\frac{2-1}{20}) + (\frac{2-1}{20}) + (\frac{2-1}{20}) + (\frac{1-1}{20}) + \frac{N_1}{20} = \frac{4}{20}.$$

Figure 5.13 shows a corresponding configuration of the probability wheel. There are only four slices of the wheel that must be assigned to E. The remaining slices are all assigned to main categories or subcategories not in E and these are labelled in Figure 5.13. ◇

We now consider the NPI lower probability for event E_2. As explained above, if $s \in A$ we have to assign N_s (see (5.27)) slices to E_2. However, E_2 contains all except a finite number of the UN main categories, so we need to think about how best to fill the slices separating different observed main categories. To avoid assigning them to E_2, we could assign these to observed main categories not in E_2 or to unobserved main categories not in E_2. There are $k - r$ observed main categories not in E_2 and there are l unobserved main categories not in E_2. Given that there are r observed main categories in E_2, this leaves $r-(k-r)-l = 2r-k-l$ separating slices to be filled.

A way in which we may be able to fill these is by using subcategories not in E_2 which belong to a main category that appears in E_2. Within each segment representing a main category described at subcategory level, $r_s - 1$ of the total number of subcategories not in E_2 are needed to separate observed subcategories in E_2. In a main category c_{j_s} with $s \notin A$, there is no upper bound on the number of unobserved subcategories not in E_2, so the separating slices to either side of the segment can always be assigned to an unobserved subcategory not in E_2 or an observed subcategory not in E_2. However, if $s \in A$, there are only $d_s + (k_{j_s} - r_s)$ subcategories that are not in E_2. Furthermore, a maximum of two subcategories per main category can be used. The number of separating slices which could be

filled in this way is

$$M_s = \begin{cases} 2 & \text{if } s \notin A \\ \min\{[d_s + (k_{j_s} - r_s) - (r_s - 1)]^+, 2\} & \text{if } s \in A \end{cases} \tag{5.29}$$

for main category c_{j_s}.

The NPI lower probability for event E_2 is therefore

$$\begin{aligned} \underline{P}(E_2) = &\sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n}) \} \\ &+ \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n}) + \frac{N_s}{n} \} \\ &+ \frac{1}{n}(2r - k - l - \sum_{s=r'+1}^{r} M_s)^+. \end{aligned} \tag{5.30}$$

Example 5.3.2. Consider the data set described in Example 5.3.1. We consider the event $Y_{21} \in \{(\text{LB} \cup \text{MB}) \cup (\text{LG} \cup \text{MG}) \cup (\text{MP}) \cup (UN_B \backslash \text{RB}) \cup (\text{LP}) \cup [UN \backslash ((UN_{Pu} \backslash \text{DPu}) \cup (\text{LO} \cup \text{MO}))]\}$. We label this event E. Let $s = 1$ correspond to B, $s = 2$ to G and $s = 3$ to P. This is an event of type E_2 (5.25), so (5.30) is used to compute the NPI lower probability for this event.

In this example, $r = 3$. The main categories for which $s \notin A$ are G and P and the only main category for which $s \in A$ is B. We have $N_1 = [(r_1 - 1) - d_1 - (k_{j_1} - r_1)]^+ = [(2 - 1) - 1 - (3 - 2)]^+ = 0$. We also have $M_1 = \min\{[d_1 + (k_{j_1} - r_1) - (r_1 - 1)]^+, 2\} = 1$, $M_2 = 2$ and $M_3 = 2$. Therefore $\sum_{s=1}^{3} M_s = 5$. The values of n_{j_s} and $n_{j_s, i_{j_s}}$ are shown in Tables 5.15 and 5.16.

	B	G	P
n_{j_s}	6	7	4

Table 5.15: Values of n_{j_s} for Example 5.3.2

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

Table 5.16: Values of $n_{j_s,i_{j_s}}$ for Example 5.3.2

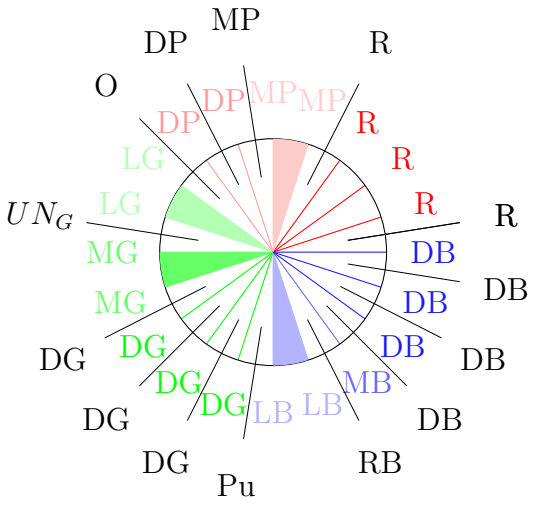

Figure 5.14: Probability wheel for Example 5.3.2

By (5.30), the NPI lower probability for the event E is

$$(\frac{2-1}{20})+(\frac{2-1}{20})+(\frac{2-1}{20})+(\frac{2-1}{20})+(\frac{1-1}{20})+\frac{N_1}{20}+\frac{(2r-k-l-5)^+}{n}=\frac{4}{20}.$$

Figure 5.14 shows a corresponding configuration of the probability wheel. There are four slices assigned to E and the remaining slices are assigned to main categories or subcategories not in E and are labelled accordingly. ◇

5.3.2 Upper probability

The NPI upper probabilities for events E_1 and E_2 are derived by assigning as many slices of the wheel as possible to the event of interest. As in Subsection 5.3.1, we know that all slices of the wheel which are between two lines representing the same main-only category or subcategory must always be assigned to that main-only category or subcategory. This means that the term

$$\sum_{s=1}^{r'}\frac{n_{j_s}-1}{n}+\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s i_{j_s}}-1}{n})$$

appears in the probability formulae for E_1 and E_2. We then consider how many separating slices we can assign to the event of interest. These include slices which separate different main categories on the wheel and slices which separate different subcategories within the same main category.

We first derive the NPI upper probability for the general event E_1 (5.24). Consider a single segment of the wheel. For a main category c_{j_s} in E_1 described at subcategory level, we have $k_{j_s} - r_s$ observed subcategories not in E_1 and r_s observed subcategories in E_1. Also, we have d_s unobserved subcategories in E_1 if $s \notin A$, or an unlimited number of UN subcategories in E_1 if $s \in A$. This means that if $s \in A$, we can always separate observed subcategories that are not in E_1 by subcategories in E_1, so all $k_{j_s} - 1$ separating slices in the segment will be assigned to E_1. In addition, we can ensure that the subcategories on the ends of the segment are in E_1. If $s \notin A$, however, we can only separate all subcategories not in E_1 if $r_s + d_s - (k_{j_s} - r_s - 1) \geq 0$, as we require at least $k_{j_s} - r_s - 1$ subcategories in E_1 to do this. Otherwise, we must assign

$$P_s = [(k_{j_s} - r_s - 1) - r_s - d_s]^+$$

separating slices to a subcategory not in E_1. In addition, if $r_s + d_s - (k_{j_s} - r_s - 1) \geq 2$, we can separate all subcategories not in E_1 and also ensure that the subcategories on the ends of the segment are both in E_1. If $r_s + d_s = k_{j_s} - r_s$, we can separate all subcategories not in E_1, but only one of the ends of the segment will be in E_1.

We now consider the wheel as a whole. There are $k - r$ observed main categories not in E_1 and r observed main categories in E_1. Of these r main categories, let $\tilde{r}$ denote the number which have both ends of their segment in E_1, i.e. the number of main categories c_{j_s} which satisfy either $s \in \{1, ..., r'\}$, $s \in A$ or the condition

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

Similarly, let r^1 denote the number of main categories which have only one end of their segment in E_1, i.e. the number of main categories such that $s \notin A$ which satisfy $r_s + d_s = k_{j_s} - r_s$, and let r^0 denote the number of main categories which

have neither end of their segment in E_1, i.e. the number of main categories such that $s \notin A$ which satisfy $r_s + d_s - (k_{j_s} - r_s - 1) \leq 0$.

There are $(k - r) + r^0 - (l + \tilde{r})$ separating slices between main categories which cannot be assigned to E_1. The NPI upper probability for event E_1 is therefore

$$\begin{aligned}\overline{P}(E_1) = &\sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n}) + \frac{k_{j_s} - 1 - P_s}{n} \} \\ &+ \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n}) + \frac{k_{j_s} - 1}{n} \} + \frac{\min\{r - r^0 + l + \tilde{r}, k\}}{n}.\end{aligned} \tag{5.31}$$

Example 5.3.3. Consider the data set described in Example 5.3.1. Suppose that we are interested in the event $Y_{21} \in \{(\text{LB} \cup \text{MB}) \cup (\text{LG} \cup \text{MG}) \cup (\text{MP}) \cup (UN_B \backslash \text{RB}) \cup (\text{LP}) \cup (UN_{Pu} \backslash \text{DPu}) \cup (\text{LO} \cup \text{MO})\}$. We label this event E. This is an event of type E_1, so (5.31) is used for the NPI upper probability for E.

In this example, $r = 3$, $l = 2$ and $k = 4$. Let $s = 1$ correspond to B, $s = 2$ to G and $s = 3$ to P. The main categories for which $s \notin A$ are G and P, and the only main category for which $s \in A$ is B. We have $P_2 = [(k_{j_2} - r_2 - 1) - r_2 - d_2]^+ = [(3 - 2 - 1) - 2 - 1]^+ = 0$ and $P_3 = [(k_{j_3} - r_3 - 1) - r_3 - d_3]^+ = [(2 - 2 - 1) - 1 - 1]^+ = 0$. The values of n_{j_s}, k_{j_s} and $n_{j_s,i_{j_s}}$ are shown in Tables 5.17 and 5.18.

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

Table 5.17: Values of n_{j_s} and k_{j_s} for Example 5.3.3

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

Table 5.18: Values of $n_{j_s,i_{j_s}}$ for Example 5.3.3

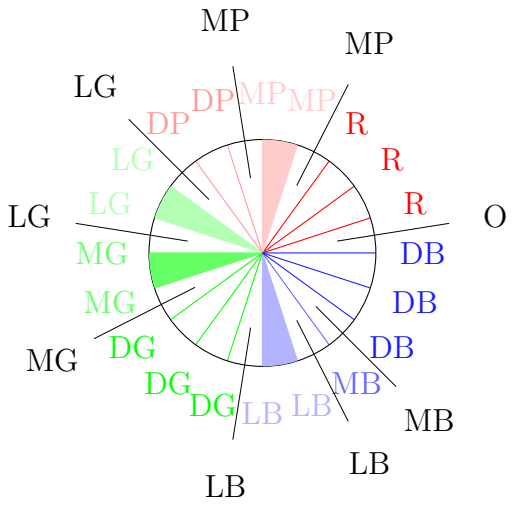

Figure 5.15: Probability wheel for Example 5.3.3

We have $r^0 = 0$ and $\tilde{r} = 3$, as both of the main categories in E for which $s \notin A$ satisfy the condition $r_s + d_s - (k_{j_s} - r_s - 1) \geq 2$. The general formula (5.31) shows that the NPI upper probability for the event E is

$$(\frac{2-1}{20}) + (\frac{2-1}{20}) + (\frac{k_{j_2} - 1 - P_2}{20}) + (\frac{2-1}{20}) + (\frac{k_{j_3} - P_3 - 1}{20})$$

$$+(\frac{2-1}{20}) + (\frac{1-1}{20}) + \frac{k_{j_1} - 1}{20} + \frac{\min\{8,4\}}{20} = \frac{13}{20}.$$

Figure 5.15 shows a corresponding configuration of the probability wheel. There are four slices of the wheel that must be assigned to E. The nine further slices that can be assigned to elements of E are labelled. ◇

We now consider event E_2 (5.25). As explained for E_1, for a single segment of the wheel representing a main category c_{j_s} in E_2 described at subcategory level, we find that if $s \in A$ we can always assign all $k_{j_s} - 1$ separating slices in the segment to E_2, whereas if $s \notin A$ we cannot necessarily do this and we must assign

$$P_s = [(k_{j_s} - r_s - 1) - r_s - d_s]^+$$

slices to a subcategory not in E_2.

Considering the wheel as a whole, we see that we can always assign all k slices separating different main categories to E_2. This is because E_2 contains all except l

of the UN main categories, and since we do not set an upper bound on the total number of unobserved categories, they can account for as many of the separating slices as we need. Note that we can also separate main categories not in E_2 using observed main categories in E_2, provided that both ends of their segments are in E_2.

This leads to the following NPI upper probability for event E_2:

$$\overline{P}(E_2) = \sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n}) + \frac{k_{j_s} - 1 - P_s}{n} \} \tag{5.32}$$

$$+ \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n}) + \frac{k_{j_s} - 1}{n} \} + \frac{k}{n}.$$

Example 5.3.4. Consider the data set described in Example 5.3.1 and suppose that we are interested in the event $Y_{21} \in \{(\text{LB} \cup \text{MB}) \cup (\text{LG} \cup \text{MG}) \cup (\text{MP}) \cup (UN_B \backslash \text{RB}) \cup (\text{LP}) \cup [UN \backslash ((UN_{Pu} \backslash \text{DPu}) \cup (\text{LO} \cup \text{MO}))]\}$. We label this event E. This is an event of type E_2, so we use (5.32) to determine the NPI upper probability for event E.

In this example, $r = 3$ and $k = 4$. Let $s = 1$ correspond to B, $s = 2$ to G and $s = 3$ to P. The main categories for which $s \notin A$ are G and P and the only main category for which $s \in A$ is B. We have $P_2 = [(k_{j_2} - r_2 - 1) - r_2 - d_2]^+ = [(3-2-1)-2-1]^+ = 0$ and $P_3 = [(k_{j_3} - r_3 - 1) - r_3 - d_3]^+ = [(2-2-1)-1-1]^+ = 0$. The values of n_{j_s}, k_{j_s} and $n_{j_s, i_{j_s}}$ are shown in Tables 5.19 and 5.20.

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

Table 5.19: Values of n_{j_s} and k_{j_s} for Example 5.3.4

By the general upper probability formula (5.32), the NPI upper probability for the

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

Table 5.20: Values of $n_{j_s,i_{j_s}}$ for Example 5.3.4

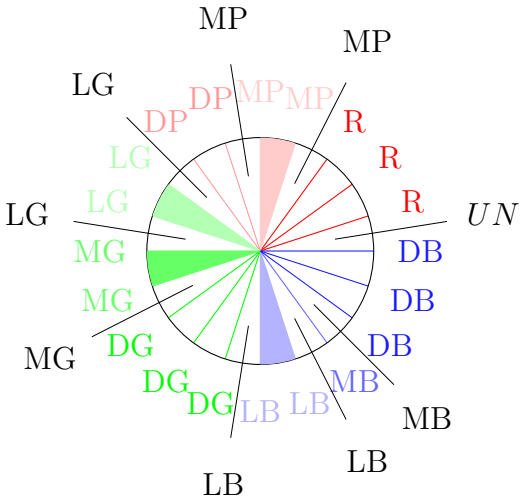

Figure 5.16: Probability wheel for Example 5.3.4

event E is

$$(\frac{2-1}{20})+(\frac{2-1}{20})+(\frac{k_{j_2}-1-P_2}{20})+(\frac{2-1}{20})+(\frac{k_{j_3}-P_3-1}{20})$$

$$+(\frac{2-1}{20})+(\frac{1-1}{20})+\frac{k_{j_1}-1}{20}+\frac{k}{20}=\frac{13}{20}.$$

Figure 5.16 shows a corresponding configuration of the probability wheel for this upper probability. ◇

It should be noted that the results of this section are closely related to the results of Section 5.1, in which the numbers of main categories and subcategories are known. The inferences in this section are either the same as the inferences in Section 5.1, or have more imprecision. In the case of more imprecision, for any event, the interval given by the NPI lower and upper probabilities derived in Section 5.1 is nested within the interval given by the NPI lower and upper probabilities derived in this section.

The properties of the results presented in this section are not considered

here in detail. It is clear, however, that the properties described in Subsections 5.2.1 to 5.2.3 also hold here. The proof of the F-probability property is under development and can be based on the proof given in Subsection 5.2.4 by writing $UN = \bigcup_{j=1}^{\infty} DN_j$ and then using a large yet finite number of these unobserved categories. Further study of the case where K and K_j, $j = 1, ..., K$, are unknown and its possible applications is an interesting topic for future research.

5.4 Classification trees with subcategory NPI

In this section we consider a method for building classification trees for subcategory data. In Chapter 4, the A-NPI-M and NPI-M algorithms were presented for finding the maximum entropy distribution consistent with the MNPI model. We now formulate a similar algorithm for approximating the maximum entropy distribution consistent with the Sub-MNPI model. An approximation is considered for the sake of computational simplicity, as is explained below. We restrict to the case where K and K_j, $j = 1, ..., K$, are known (see Section 5.1).

In Chapter 4, it was proven that for data with main categories only, the NPI lower and upper probabilities for a general event can always be derived from the singleton probabilities L_j and U_j (see Theorem 4.1.1). Let L_{j,i_j} and U_{j,i_j} denote the NPI lower and upper probabilities for the event that the next observation is in subcategory s_{j,i_j}. Then the equivalent statement to Theorem 4.1.1 for the Sub-MNPI model would be that $\underline{P}(E)$ is given by the expression

$$\max\{\sum_{j\in J} L_j + \sum_{j\in J^*}\sum_{i_j\in I_j} L_{j,i_j}, 1 - \sum_{j\in F} U_j - \sum_{j\in J^*}\sum_{i_j\in \overline{I_j}} U_{j,i_j}\} \tag{5.33}$$

and $\overline{P}(E)$ is given by the expression

$$\min\{\sum_{j\in J} U_j + \sum_{j\in J^*}\sum_{i_j\in I_j} U_{j,i_j}, 1 - \sum_{j\in F} L_j - \sum_{j\in J^*}\sum_{i_j\in \overline{I_j}} L_{j,i_j}\} \tag{5.34}$$

However, this is not the case. For subcategory data, the NPI lower and upper probabilities for a general event cannot be determined from the singleton probabilities. Because of the need to configure the wheel such that all subcategories

within the same main category are grouped together, the set of probability distributions that are consistent with the Sub-MNPI model is more restricted than that of the original MNPI model and the interval probability for event E is often narrower than the interval given by the expressions (5.33) and (5.34). The interval consisting of the NPI lower probability (5.7) and the NPI upper probability (5.12) is therefore not an F-probability interval in the sense of Weichselberger. This is illustrated in Example 5.4.1.

Example 5.4.1. Consider a multinomial data set with possible main categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). Observations in B are further classified as light blue (LB), medium blue (MB) or dark blue (DB), observations in Y are further classified as light yellow (LY), medium yellow (MY) or dark yellow (DY) and observations in G are further classified as light green (LG), medium green (MG), dark green (DG) or other green (OG). The data set consists of twenty observations altogether, including 1 LB, 1 MB, 2 DB, 1 LG, 1 MG, 1 DG, 1 OG, 1 LY, 1 MY, 1 DY, 2 O, 3 R and 4 P.

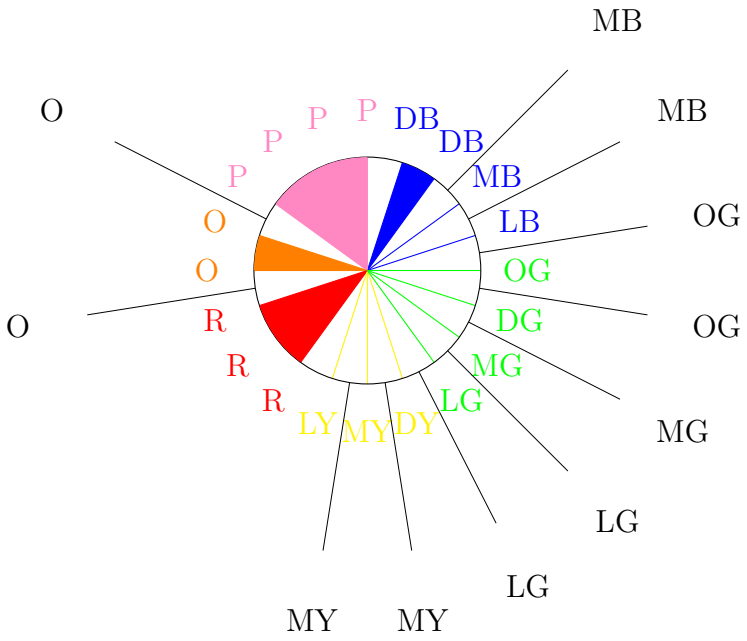

Figure 5.17: Probability wheel for Example 5.4.1

Suppose that we are interested in the event $Y_9 \in \{\text{R,P,LY,DY,DG,LB,DB}\}$. We label this event E. According to (5.7), the NPI lower probability for the event E is

equal to $\frac{8}{20}$. There is no configuration of the wheel which results in a smaller lower probability. A possible configuration of the wheel which gives the lower probability $\frac{8}{20}$ is shown in Figure 5.17. However, (5.33) gives

$$\max\{\sum_{j\in J} L_j + \sum_{j\in J^*}\sum_{i_j\in I_j} L_{j,i_j}, 1 - \sum_{j\in F} U_j - \sum_{j\in J^*}\sum_{i_j\in \overline{I_j}} U_{j,i_j}\} = \max\{\frac{6}{20}, \frac{7}{20}\} = \frac{7}{20}.$$

This discrepancy is due to the extra restriction on the configuration of the wheel which arises from the fact that we group together all subcategories within the same main category. ◇

We now construct an algorithm which approximates the maximum entropy distribution consistent with the Sub-MNPI model. The process of computing the distribution is carried out in two stages. Initially, we consider observations at main category level only. We apply the NPI-M algorithm described in Section 4.2. This gives a maximum entropy probability $p_{maxE}(c_j)$ for each main category. We then need to share this probability mass between all subcategories of c_j. In order to do this in such a way that the resulting distribution would always correspond to a valid configuration of the probability wheel, it would be necessary to specify the exact wheel configuration for the results of the NPI-M algorithm and to then consider each category individually. This is because the main category segments do not have a fixed configuration. A segment may consist of a whole number of slices or may sometimes contain a fraction of a slice. There may be a fraction of a slice at just one end of the segment or at both ends of the segment. This problem is clarified further by Example 5.4.2.

In order to give a computationally straightforward algorithm, we use an approximation at this stage of the process. This approximation is analogous to that of the A-NPI-M algorithm presented in Section 4.1. We share the probability mass $p_{maxE}(c_j)$ as evenly as possible between the subcategories, in such a way that the probability $\widehat{p}_{j,i_j}$ that is assigned by the algorithm to subcategory s_{j,i_j} is within the interval $[L_{j,i_j}, U_{j,i_j}]$. However, in practice there may not be a configuration of the wheel that corresponds to the resulting distribution $\widehat{p}$. This is

illustrated in Example 5.4.2.

Let $K(i)_j$ represent the number of subcategories in main category c_j that have been observed i times. Suppose that we have already applied the NPI-M algorithm and we have the results $p_j = p_{maxE}(c_j), j = 1, ..., K$. This means that for each main category c_j, we have a segment consisting of np_j slices. Of these slices, $n(\sum_{i=1}^{K_j} L_{j,i_j})$ must be assigned to observed subcategories in c_j. We therefore have probability mass $p_j - \sum_{i=1}^{K_j} L_{j,i_j}$ that may be assigned to any available subcategory in c_j and this is termed remaining probability mass. For each c_j, we share this remaining probability mass between subcategories of c_j, beginning with subcategories with the fewest observations. This leads to the algorithm shown below. The algorithm is programmed in Weka software for possible use in practical applications, but is written here in pseudo-code.

Sub-A-NPI-M

For $j = 1$ to K and for $i = 1$ to K_j

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

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

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

$t \leftarrow 0$;

While $(rem > 0)$ do

If $(\frac{1}{n}(K(t)_j + K(t+1)_j) < rem)$

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

$rem \leftarrow rem - \frac{1}{n}$;

Else

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

$rem \leftarrow 0$;

$t \leftarrow t + 1$;

The Sub-A-NPI-M algorithm is illustrated in Example 5.4.2.

Example 5.4.2. Consider a multinomial data set with observed main categories blue (B), green (G), red (R) and pink (P) and unobserved main category orange (O). Observations in B are further classified as light blue (LB) or dark blue (DB) and observations in G are further classified as light green (LG) or dark green (DG). The data set consists of twenty observations altogether, including 5 DB, 5 DG, 5 R and 5 P.

First, considering the data at main category level only, we apply the NPI-M algorithm (see Section 4.2). Here $K(0) < K'$, so we use the algorithm described in Subsection 4.2.2. The NPI-M algorithm initially assigns the lower probability $\frac{n_j-1}{n}$ to categories R, B, G and P. So $p(\mathrm{R}) = p(\mathrm{B}) = p(\mathrm{G}) = p(\mathrm{P}) = \frac{4}{20}$. The algorithm assigns probability $\frac{1}{20}$ to the unobserved category O.

For $i = 1, ..., 3$, $K(i) + K(i+1) = 0$. Taking $i = 4$, $K(4) + K(5) = 4$ and $mass = 3$, therefore $K(i) + K(i+1) > mass$ and

$$W = \min\{mass + 1 + K(4), K(4) + K(5)\} = 4.$$

This means that the four categories observed five times are assigned the probability

$$p_{maxE}(c_j) = \frac{4}{20} + \frac{mass}{20(\min\{mass + 1 + K(4), K(4) + K(5)\})} = \frac{4}{20} + \frac{3}{4 \times 20} = \frac{19}{80}.$$

So the maximum entropy probabilities assigned to the main categories {O,R,B,G,P} are

$$\{\frac{1}{20}, \frac{19}{80}, \frac{19}{80}, \frac{19}{80}, \frac{19}{80}\}. \tag{5.35}$$

A configuration of the wheel corresponding to this distribution is shown in Figure 5.18. The separating slices are shared in such a way that B, R, G and P are each assigned $\frac{3}{4}$ of a separating slice. We now consider the subcategories. The maximum entropy probabilities for the main categories (5.35) are distributed over the subcategories using the Sub-A-NPI-M algorithm. For main category B we have $\underline{P}(DB) = \frac{4}{20}$ and $\underline{P}(LB) = 0$. For main category G we have $\underline{P}(DG) = \frac{4}{20}$ and

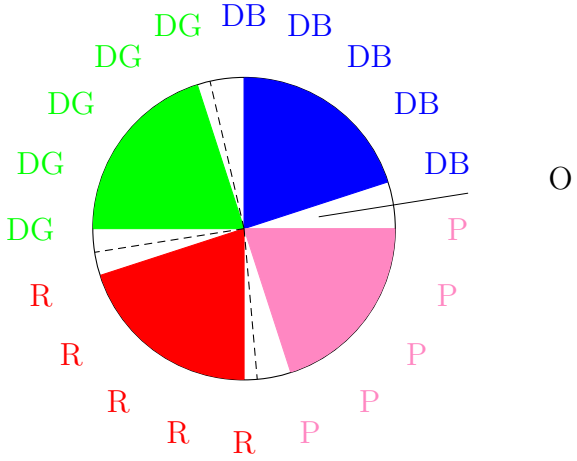

Figure 5.18: Probability wheel for Example 5.4.2

$\underline{P}(LG) = 0$. Applying the Sub-A-NPI-M algorithm, we find that

$$rem = \frac{19}{80} - \frac{4}{20} = \frac{3}{80}$$

for both of these main categories. Taking $t = 0$,

$$\widehat{p}(LB) = 0 + \min\{\frac{rem}{K(t)_j + K(t+1)_j}, \frac{1}{n}\} = 0 + \min\{\frac{3}{80}, \frac{1}{20}\} = \frac{3}{80}$$

and

$$\widehat{p}(LG) = 0 + \min\{\frac{rem}{K(t)_j + K(t+1)_j}, \frac{1}{n}\} = 0 + \min\{\frac{3}{80}, \frac{1}{20}\} = \frac{3}{80}.$$

So the probabilities assigned to the set of subcategories {LB,DB} are

$$\{\frac{3}{80}, \frac{4}{20}\}$$

and the probabilities assigned to the set of subcategories {LG,DG} are

$$\{\frac{3}{80}, \frac{4}{20}\}.$$

In the configuration shown above, this distribution can be achieved for the B subcategories in a way that maintains a valid arrangement of the wheel. This is because the remaining probability mass is all contained in one portion, located at one end of the B segment. However, for the G subcategories there is not a valid arrangement of the wheel that results in this distribution. This is because the remaining probability mass is split up and some is located at either end of the G segment. ◇

The Sub-A-NPI-M algorithm can be implemented for building classification trees in the same way that the A-NPI-M and NPI-M algorithms were used in Chapter 4. We have not yet considered this further, but it is of interest to explore this topic in future research.

5.5 Concluding remarks

In this chapter we presented the Sub-MNPI model for inferences from multinomial data described at subcategory level as well as at main category level. NPI lower and upper probabilities were derived for the general events of interest and some fundamental properties of these inferences were proven. An algorithm was then presented for approximating the maximum entropy distribution consistent with these inferences. Further study of the Sub-MNPI model would be an interesting future research topic and the work presented here could be extended by further study of the properties of the results derived in Section 5.3 where K and K_j, $j = 1, ..., K$, are unknown. The application to classification could be investigated further and it would be of particular interest to compare classification trees built using the Sub-A-NPI-M algorithm presented in Section 5.4 with classification trees constructed by ignoring the hierarchical relationship between the categories and subcategories and simply using the NPI-M algorithm presented in Section 4.2. Further research to develop an algorithm for the true maximum entropy distribution consistent with the Sub-MNPI model would also be a useful extension of this work. In the future, other applications of NPI for subcategory data could be investigated and the Sub-MNPI model could also be developed further by considering inferences about multiple future observations and by introducing further layers, e.g. subsubcategories, to the hierarchy. Such developments would be of theoretical and practical interest.

Chapter 6

Conclusion

In this thesis, several extensions to the theory of NPI for multinomial data were presented, alongside applications of this theory in the areas of selection and classification. The research presented here comprises a substantial contribution to the development of NPI as a useful and versatile inferential framework and as one of the fastest-growing areas within the field of statistical inference using imprecise probability.

In Chapter 3, an extension of the MNPI model was presented which enabled inferences about multiple future observations. These inferences were applied to the problems of category selection and subset selection for multinomial data. The results of this chapter could be extended further still: the derivation of lower and upper probabilities for all events of interest involving multiple future observations is an important subject for future research and would be of use in more complex selection problems as well as in other areas of application.

In Chapter 4, algorithms based on the MNPI model were developed for building classification trees. These were tested on forty data sets and it was shown that NPI-based classifiers performed well compared with other classifiers from the literature. The application of NPI to classification provides a number of opportunities for future research. Further study of NPI-based classification trees could be pursued, including a more in-depth analysis of a larger number of data

sets and a more detailed examination of split variable selection bias. Aside from classification trees, the application of NPI to other types of classification could be considered. It is also important to investigate imprecise classification with NPI, where the classifier output is a set of categories as opposed to a single category.

In Chapter 5, the Sub-MNPI model was presented for inferences about a future observation from a set of multinomial data with subcategories. NPI lower and upper probabilities were derived for all events of interest and some fundamental properties of the model were proven. Also, an algorithm based on the Sub-MNPI model was proposed for building classification trees. The Sub-MNPI model is an important contribution to the theory of NPI and in future it could be extended further to enable inferences about multiple future observations. Applications of the Sub-MNPI model to classification and to other areas could also be investigated further.

The increasing popularity of NPI and other imprecise probability theories has led to widespread interest in applications of NPI to practical problems. The continued investigation of such applications is an important future research topic, but there is a need for further extensions to the MNPI model in order to analyse multinomial data sets in a meaningful way. In addition to the inferences about multiple future observations discussed above, there are a number of issues to consider. It is important to incorporate ordinal data into the NPI methodology, as categories often have a natural ordering to them, and a treatment of ordinal data distinct from categorical data could give interesting results. Also, the idea of subcategory data should be taken further and the introduction of further layers e.g. subsubcategories to form a nested hierarchy should be considered. Techniques for handling missing data should also be assimilated, as missing values are a common occurrence in data sets. These extensions to the MNPI model would be of interest theoretically and may also lead to exciting new applications of NPI in the future.

Bibliography

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

[2] Abellan, J. and Moral, S. (2003) Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **11**, 587-597.

[3] Abellan, J., Baker, R.M. and Coolen, F.P.A. (2009) Maximising entropy on the nonparametric predictive inference model for multinomial data. Submitted.

[4] Asuncion, A. and Newman, D.J. (2007) *UCI Machine Learning Repository [http://www.ics.uci.edu/mlearn/MLRepository.html].* University of California, Irvine, School of Information and Computer Science.

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

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

[7] Baker, R.M. and Coolen, F.P.A. (2009) Category selection for multinomial data. *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, 21-30.

[8] Baker, R.M. and Coolen, F.P.A. (2009) Nonparametric predictive category selection for multinomial data. Submitted.

[9] Bechhofer, R.E. (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, **25**, 16-39.

[10] Bechhofer, R.E., Santner, T.J. and Goldsman, D.M. (1995) *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons.* Wiley.

[11] Bernard, J.M. (2009) Special issue on the Imprecise Dirichlet Model. *International Journal of Approximate Reasoning*, **50**, 201-268.

[12] Boole, G. (1854) *An Investigation of the Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities.* Walton and Maberley, London.

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

[14] Coolen, F.P.A. and Augustin, T. (2005) Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. *Proceedings of the Fourth International Symposium on Imprecise Probability: Theories and Applications*, 125-134.

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

[16] Coolen, F.P.A. (2010) *International Encyclopaedia of Statistical Sciences* (Nonparametric Predictive Inference chapter). Springer.

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

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

[19] Coolen, F.P.A. and Coolen-Schrijner, P. (2006) Nonparametric predictive subset selection for proportions. *Statistics and Probability Letters*, **76**, 1675-1684.

[20] Coolen, F.P.A. and Coolen-Schrijner, P. (2007) Nonparametric predictive comparison of proportions. *Journal of Statistical Planning and Inference*, **137**, 23-33.

[21] 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**, 259-277.

[22] 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.

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

[24] De Campos, L.M., Huete, J.F. and Moral, S. (1994) Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **2**, 167-196.

[25] Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis.* Wiley, New York.

[26] Gupta, S.S. (1965) On some multiple decision (selection and ranking) rules. *Technometrics*, **7**, 225-245.

[27] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutermann, B. and Witten, I.H. (2009) The Weka data mining software: an update. *SIGKDD Explorations*, **11**, 10-18.

[28] Hampel, F. (2009) Nonadditive probabilities in statistics. *Journal of Statistical Theory and Practice*, **3**, 11-23.

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

[30] Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, **2**, 1137-1143.

[31] Levi, I. (1980) *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability and Chance.* MIT Press.

[32] Maturi, T., Coolen-Schrijner, P. and Coolen, F.P.A. (2009) Nonparametric predictive selection with early experiment termination. Submitted.

[33] Miller, G. (1955) Note on the bias of information estimates. *Information Theory in Psychology*, 95-100.

[34] Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning*, **1** 81-106.

[35] Quinlan, J.R. (1993) *Programs for Machine Learning*. Morgan Kaufmann.

[36] Rice, J.A. (1995) *Mathematical Statistics and Data Analysis*. Wadsworth, California.

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

[38] Shannon, C.E. (1948) A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379-423.

[39] Strobl, C. (2005) Variable selection in classification trees based on imprecise probabilities. *Proceedings of the Fourth International Symposium on Imprecise Probability: Theories and Applications.*

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

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

[42] Wasserman, L. and Kadane, J.B. (1996) *Bayesian Analysis in Statistics and Econometrics.* Wiley, New York.

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

[44] Weichselberger, K. (2001) *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung.* Physica-Verlag Heidelberg.

[45] Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

[46] Zaffalon, M. (1999) A credal approach to naive classification. *Proceedings of the First International Symposium on Imprecise Probability: Theories and Applications*.

[47] Zaffalon, M. (2001) Statistical inference of the naive credal classifier. *Proceedings of the Second International Symposium on Imprecise Probability: Theories and Applications*.

[48] Zaffalon, M. (2002) The naive credal classifier. *Journal of Statistical Planning and Inference*, **105**, 5-21.