Category Selection for Multinomial Data

Rebecca M. Baker
Department of Mathematical Sciences
University of Durham, England
r.m.baker@durham.ac.uk

Frank P.A. Coolen
Department of Mathematical Sciences
University of Durham, England
frank.coolen@durham.ac.uk

Abstract

A new method is presented for selecting a single category or the smallest subset of categories, based on observations from a multinomial data set, where the selection criterion is a minimally required lower probability that (at least) a specific number of future observations will belong to that category or subset of categories. The inferences about the future observations are made using an extension of Coolen and Augustin's nonparametric predictive inference (NPI) model to a situation with multiple future observations.

Keywords. imprecise probability, predictive inference, categorical data, selection

1 Introduction

Selection is a wide-ranging topic in statistics for choosing the optimal member(s) of some group. This group may be, for example, a set of data 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. Existing methods for this type of selection [2] are all non-predictive, i.e. the selection of the optimal category is based solely on hypothesis testing and does not use any type of predictive inference.

NPI for learning from multinomial data in the absence of prior knowledge has been developed by Coolen and Augustin [1, 6, 7]. The model gives predictive inferences about a single future observation in the form of probability intervals $P = [\underline{P}, \overline{P}]$. Throughout this paper, P denotes interval probability, which we often just call 'probability'. When an explicitly precise probability is used, it is denoted by p. NPI is based on a probability wheel representation of the data, where each category is represented by a segment of the wheel.

Selection methods based on NPI have been developed by Coolen and van der Laan [3] and Coolen and Coolen-Schrijner [4, 5]. These methods use predictive inferences which are based on past observations, and make use of Hill's assumption A_n [9].

Coolen and van der Laan [3] developed an NPI selection method for real-valued data from k different sources. Their objective was to select the source which would provide the largest next observation. Probabilities were determined for the event that the next observation from one source would 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 would contain the source providing the largest next observation, and second, they found the interval probability that the next observations from every source in some subset would all exceed the next observations from the remaining sources.

Coolen and Coolen-Schrijner [4, 5] developed an NPI selection method for Bernoulli data from k different groups. Their objective was to select the group which would have the highest number of future successes. Here, inferences were made about m future observations rather than just the next observation. Subsets of the groups were also considered [4], and probabilities were 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.

In this paper, we discuss the use of NPI for selection from a multinomial data set. We consider selection of a single optimal category, and selection of an optimal subset of categories, where we define the optimal subset to be the subset which satisfies the required probability criterion, is of minimal size and has the largest lower probability amongst all subsets of the same size.

2 Predictive category selection

We develop NPI for category selection from a multinomial data set. We have K possible categories, labelled $c_1, ..., c_K$, and our aim is to select the category with the largest probability of occurrence. Suppose that we have a data set consisting of n observations, and let $n_1, ..., n_K$ denote the number of observations in categories $c_1, ..., c_K$ respectively. We consider m future observations, and select a category based on predictive inferences about these m observations. These inferences will be made by using and adapting the general theory of nonparametric predictive inference for multinomial data [1, 6, 7], discussed previously. Let the vector of random quantities $(M_1, ..., M_K)$ denote the number of the m future observations that belong to categories $c_1, ..., c_K$, such that $\sum_{j=1}^{K} M_j = m$.

2.1 One future observation

The simplest case is where $m = 1$, so inference is about one future observation. We may want to select a single category with the largest probability of occurrence. According to the NPI model [7], the lower and upper probabilities that the future observation will belong to category c_j are

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

where $(x)^+$ denotes $\max\{x, 0\}$, and

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

The above formulae are derived through the use of the probability wheel model [6], as illustrated in the example below. We can evaluate these probabilities for each of the possible categories and then select the category with the highest probability.

Example 2.1. *Suppose that our possible categories are blue (B), red (R), yellow (Y) and green (G). Our data set consists of 8 observations: 3 B, 2 G, 2 Y and 1 R. We want to select a single category with the highest probability that the next observation will be in that category. First, we find the probability that the next observation will be blue. Let n_B denote the number of B observations in the data set, and let M_B denote the number of future B observations. The minimum number of slices of the wheel that we can assign to B is equal to $n_B - 1 = 3 - 1 = 2$. This leads to the lower probability $\underline{P}(M_B = 1) = \frac{n_j - 1}{n} = \frac{2}{8}$.*

The maximum number of slices of the wheel that we can assign to B is equal to $n_B + 1 = 3 + 1 = 4$. This leads to the upper probability $\overline{P}(M_B = 1) = \frac{n_j + 1}{n} = \frac{4}{8}$.

We then carry out the same process for the other categories, and we find that $P(M_Y = 1) = P(M_G = 1) = [\frac{1}{8}, \frac{3}{8}]$, and $P(M_R = 1) = [0, \frac{2}{8}]$. So we select the blue category.

Theorem 2.1. *When $m = 1$, and we want to select a single category with the largest probability of occurrence, it is always optimal to choose the category which has the greatest number of observations in the data set.*

Proof. We select the category with the highest probability $P(M_j = 1)$, where $P(M_j = 1) = [\frac{n_j - 1}{n}, \frac{n_j + 1}{n}]$, so it is optimal to select the category with the largest value of n_j. □

2.2 Multiple future observations

Whereas Coolen and Augustin [6, 7] only considered one future observation, we now consider inferences about multiple future observations, so $m > 1$. Suppose that our data set is represented on a probability wheel, and the n slices on the wheel are numbered 1 to n. Each of our m future observations must fall into one of these n slices. Let the vector $(S_1, ..., S_n)$ denote the number of future observations which fall into slices 1 to n, respectively. The total number of different arrangements of these m observations is $\binom{n+m-1}{m}$ [8], which leads to 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 $S + 1$ observations is equal to

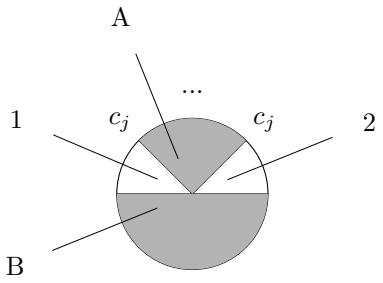
$$\binom{(S - 1) + f}{f}. \quad (1)$$

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

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

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

We can use NPI to find the probabilities that precisely m_j of the m future observations will belong to category c_j. The bounds derived here are the most conservative bounds achievable within the NPI framework, due to the way in which the slices of the wheel are assigned to categories. This is explained below. The diagram illustrates the relevant segments of the wheel.



It is assumed throughout this section that $1 < n_j < n - 1$. In the case $n_j \leq 1$, we are not forced to assign any slices of the wheel to c_j, leading to $\underline{P}(M_j = m_j) = 0$.

The shaded segment A represents all slices which must be assigned to c_j. There are $n_j - 1$ such slices. By (1), the number of different arrangements of m_j future observations within this segment is $\binom{n_j - 2 + m_j}{m_j}$.

The shaded segment B represents all slices which must be assigned to a category other than c_j. There are $n - n_j - 1$ such slices. By (1), the number of different arrangements of $m - m_j$ future observations within this segment is $\binom{n - n_j - 2 + (m - m_j)}{m - m_j}$.

Multiplying these two binomial coefficients gives us the minimum number of arrangements in which m_j future observations are in c_j, showing that the lower probability is equal to

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

This general formula is applicable to any positive integers m and m_j such that $m_j \leq m$.

We can also find the equivalent upper probability. We now want to maximise the number of arrangements of the m future observations in which m_j future observations are in c_j. There are $n_j + 1$ slices of the wheel which we can allocate to category c_j, including two slices which we may or may not assign to c_j, which we will term 'optional slices' (labelled 1 and 2 in the diagram above).

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. 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. 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 count any arrangement 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 arrangements, as the observation could fall either in slice 1 or in slice 2. By similar reasoning, for $T = 2$, there are three possible arrangements. In general, there are $T + 1$ possible arrangements for each value of T.

However, there are a number of different orderings that give T observations in the optional slices. Let X be a non-negative 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 with one or more observations in the optional slices is equal to

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

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 upper probability

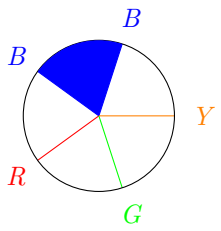
$$\overline{P}(M_j = m_j) = \binom{n + m - 1}{m}^{-1} [\binom{n_j - 2 + m_j}{m_j} \times \binom{n - n_j - 2 + (m - m_j)}{m - m_j} + \sum_{T=1}^{m} \sum_{X=\{T-(m-m_j)\}^+}^{\min\{m_j, T\}} \times (T+1) \binom{n_j - 2 + (m_j - X)}{m_j - X} \times \binom{n - n_j - 2 + (m - m_j) - (T - X)}{m - m_j - (T - X)}]. \quad (3)$$

Again, this formula holds for any positive integers m and m_j such that $m_j \leq m$. As before, it is assumed here that $n_j \geq 2$. An unobserved category can be assigned at most one slice of the wheel, leading to $\overline{P}(M_j = m_j) = \binom{n+m-1}{m}^{-1}\binom{n-n_j-2+m-m_j}{m-m_j}$. In the case $n_j = 1$, the formula reduces to

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

In the case $n_j \geq n-1$, every slice on the wheel may be assigned to category j and furthermore there is only one optional slice.

Example 2.2. *Suppose that our possible categories are blue (B), red (R), yellow (Y) and green (G). Our data set consists of 5 observations as shown on the probability wheel below.*



We want to make inferences about 3 future observations, and we want to find the probability that precisely two of these are blue. To find the lower probability, we use (2) with $m_B = 2$. Using the values $n = 5$, $m = 3$ and $n_j = 2$, this gives

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

To find the upper probability, we use (3) with $m_B = 2$. This gives

$$\overline{P}(M_B = 2) = \frac{1}{35}[2+2+4+3+6+4] = \frac{21}{35}.$$

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

Theorem 2.2. *For general m, when selecting the category which has the largest lower or upper probability of containing all of the future observations, it is optimal to select the category with the greatest number of observations.*

Proof. The general formulae for the lower probability (2) and upper probability (3) 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 $T - X = 0$. We find that

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

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 these lower and upper probability formulae are both increasing in n_j, it is always optimal to select the category with the largest value of n_j, ie. the greatest number of data observations. □

It is also of interest to investigate which value of n_j will maximise the lower probability $\underline{P}(M_j = m_j)$. We will henceforth call this value n_j^*. Plotting $\underline{P}(M_j = m_j)$ against values of n_j ranging from 1 to n shows the graph to be monomodal with a smooth line of best fit. Intuitively, we expect that the 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 that are in c_j. We will now formally assess which value of n_j gives the maximal lower probability.

Theorem 2.3. *For general m, the value of n_j which will maximise $\underline{P}(M_j = m_j)$ is the integer which lies in the interval $[1 + \frac{m_j}{m}(n-3), 2 + \frac{m_j}{m}(n-3)]$.*

Proof. The proof follows from considering the two ratios

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

and

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

□

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

$$1 + \frac{m_j}{m}(n-3) \leq \frac{nm_j}{m} \iff m_j \geq \frac{1}{3}m$$

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

We see that if $\frac{1}{3}m \leq m_j \leq \frac{2}{3}m$, then $\frac{nm_j}{m}$ will indeed be within the interval. We can also show 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.

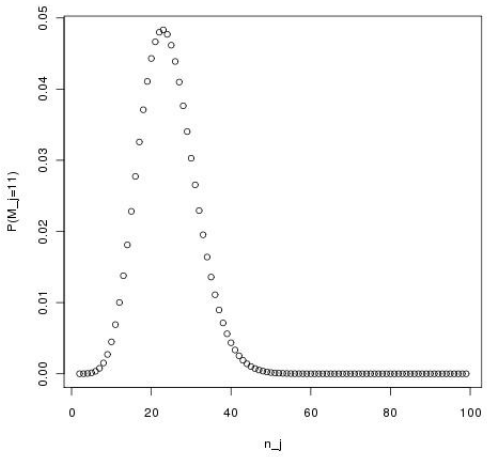
Corollary 2.1. *For general m, when selecting a category which maximises $\underline{P}(M_j = m_j)$, the optimal category is selected as follows:*

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 lower probability is optimal.*

We also notice that if we have a lot of observations and if both m_j and m are very large, then $\frac{m_j}{m}$ will tend to some limit l and therefore 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 optimal value of the ratio will tend to the same limit l, as is to be expected.

Example 2.3. *Suppose we have a categorical data set consisting of 100 observations. There are 4 possible categories: blue (B), red (R), yellow (Y) and green (G). We have observed 20 B, 25 R, 28 Y and 27 G. We are making inferences about the next 50 observations, and we wish to select the category that maximises the lower probability $\underline{P}(M_j = 11)$.*

The plot of $\underline{P}(M_j = 11)$ against all possible values of n_j is shown below. From this graph, we expect that n_j^* will be between 20 and 25, as this is where the peak occurs.



By Theorem 2.3, 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, and so by Corollary 2.1 we must look at either side of the interval.

To the left of the interval, we have $n_j = 20$ corresponding to the blue category. By (2), 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 the red category. The lower probability here is $\underline{P}(M_R = 11) = 0.0462$. As the second probability is largest, we see that $n_j = 25$ is the optimal choice, and so we select red as our optimal category.

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

The other event of interest here is that *at least* m_j of the m future observations will belong to category c_j. For the lower probability, we again count the minimum number of relevant arrangements of the future observations. However, we are now interested in all arrangements which have R future observations which 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 $\binom{n_j-2+R}{R}$ different arrangements within this segment. We must also consider the remaining $m - R$ observations. Contrary to our lower probability formula above (2), arrangements with one or more observations in an optional slice will now be counted. We did not count these when finding the lower probability $\underline{P}(M_j = m_j)$, because for example an arrangement with m_j observations in segment A and 1 in an optional slice could be allocated to the event $M_j = m_j + 1$ when deriving $\underline{P}(M_j = m_j)$. However, such arrangements are now relevant because we are simultaneously considering all events $M_j \in \{m_j, m_j + 1, ..., m\}$.

By (1), the number of different arrangements of $m - R$ future observations within the shaded segment B plus the two optional slices is equal to $\binom{n-n_j+(m-R)}{m-R}$.

Multiplying the two binomial coefficients above leads to the minimum number of arrangements in which R future observations are in c_j. We now sum over R from m_j to m, which gives the lower probability

$$\underline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1} \sum_{R=m_j}^{m} \binom{n_j-2+R}{R} \times \binom{n-n_j+(m-R)}{m-R}. \quad (4)$$

It is assumed here that $n_j \geq 2$, because otherwise the lower probability will be zero. We also assume $m_j > 0$.

To find the corresponding upper probability, we have to 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 the arrangements described above, so all of the $\binom{n_j-2+R}{R}\binom{n-n_j+(m-R)}{m-R}$ arrangements will be included in our total, where $m_j \leq R \leq m$. However, we also want to include any 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 arrangements of these observations for each value of T. Therefore, by (1), the number of different arrangements is equal to

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

Summing together both of the above numbers gives the total number of relevant arrangements, leading to the upper probability

$$\begin{aligned}\overline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1} [&\sum_{R=m_j}^{m} \binom{n_j-2+R}{R} \times \binom{n-n_j+(m-R)}{m-R} \\ &+ \sum_{Y=0}^{m_j-1} \sum_{T=m_j-Y}^{m-Y} (T+1)\binom{n_j-2+Y}{Y} \\ &\times \binom{n-n_j-2+(m-Y-T)}{m-Y-T}].\end{aligned} \tag{5}$$

As before, we assume $n_j \geq 2$ and $m_j > 0$. In the cases $n_j = 1$ and $n_j = 0$, the formula reduces to

$$\overline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1} \sum_{T=m_j}^{m} (T+1) \times \binom{n-n_j-2+(m-T)}{m-T}$$

and

$$\overline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1} \sum_{T=m_j}^{m} \times \binom{n-n_j-2+(m-T)}{m-T}$$

respectively.

These formulae can be used in a number of different ways. For example, suppose we wanted to select a category for which there was at least a 75% lower probability that two or more of the future observations would be in that category. We would use the above formulae to find all c_j such that $\underline{P}(m_j \geq 2) \geq 0.75$. Alternatively, suppose we wanted to select the category which was most likely to contain 10% or more of the future observations. We would evaluate $P(m_j \geq \frac{m}{10})$ for each of the possible categories, and then select the category according to these values.

This method of selection is illustrated in the example below.

Example 2.4. *Consider Example 2.2, where our possible categories are blue (B), red (R), yellow (Y) and green (G) and our data set consists of 5 observations as shown on the probability wheel in Example 2.2.*

We are making inferences about 3 future observations, and we want to select the category with the highest probability of containing at least one third of the future observations. To find the lower probability of the event $M_j \geq \frac{m}{3}$, we use (4) with $m_j = 1$. We first consider the blue category. 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 upper probability, we use (5) with $m_j = 1$. For blue, 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 $P(M_B \geq 1) = [\frac{15}{35}, \frac{31}{35}]$. We investigate the three remaining categories in the same way, and we find that $P(M_j \geq 1) = [0, \frac{25}{35}]$ for all three categories. So the category we select here is blue.

3 Predictive subset selection

We now consider the use of predictive methods to select a subset of categories, rather than a single category, from a multinomial data set. As before, we have K possible categories, and we have a data set consisting of n observations where $n_1, ..., n_K$ denote the number of times we have observed categories $c_1, ..., c_K$ respectively. Recall that k represents the total number of categories that have been observed. We will select our subset based on inferences about m future observations. Our inferences use the general theory of nonparametric predictive inference [7].

3.1 One future observation

In this case, our aim will be to select a subset in order to maximise the NPI lower probability that the next observation, Y_{n+1}, belongs to a category within that subset.

Let S denote our selected subset of categories. Let OS denote the index set for already-observed categories in S, and let US denote the index set for unobserved categories in S. The sizes of these sets are denoted r and l respectively. Then, according to the NPI model [7], the formula for the lower probability $\underline{P}(Y_{n+1} \in S)$ is

$$\underline{P}(Y_{n+1} \in S) = \sum_{j \in OS} \frac{n_j - 1}{n} + \frac{(2r + l - K)^+}{n} \quad (6)$$

and the formula for the upper probability $\overline{P}(Y_{n+1} \in S)$ is

$$\overline{P}(Y_{n+1} \in S) = \sum_{j \in OS} \frac{n_j - 1}{n} + \frac{\min\{2r + l, k\}}{n}. \quad (7)$$

Our objective is to find some 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 the one with maximum lower probability.

Example 3.1. *Consider Example 2.1, where our possible categories are blue (B), red (R), yellow (Y) and green (G), and our data set consists of 8 observations including 3 B, 2 G, 2 Y and 1 R. Now, 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 2.1, 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 will not satisfy our requirements. We instead look for a subset of size 2. Consider the subset* $S = \{B, G\}$. *Here,* $r = 2$ *and* $l = 0$. *The formula (6) gives*

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

This satisfies the selection criterion. Applying the same formula to other possible subsets of size 2 shows that $\frac{3}{8}$ *is the highest lower probability that we can achieve with a subset of size 2. So the subset we select is* $S = \{B, G\}$.

Theorem 3.1. *When* $m = 1$, *and we want to select a subset of categories according to our aforementioned definition of the optimal subset, it is always optimal to add categories to the subset in decreasing order of number of observations in the data set.*

Proof. We select a subset according to which gives the highest lower probability $\underline{P}(Y_{n+1} \in S)$. The addition of an already-observed category to S will add $\frac{n_j - 1}{n}$ to the first term in the lower probability formula and will add 2 to the second term. The addition of an unobserved category to S will add 0 to the first term and 1 to the second term. So we should always add observed categories before unobserved categories. Furthermore, the observed categories which will give the largest increase to the lower probability when added to S are those with the largest values of n_j. So it is always optimal to include categories in S in decreasing order of n_j, ie. in decreasing order of the number of observations. □

3.2 m future observations

We now consider inferences about multiple future observations. This requires some new notation: let M_S represent the number of future observations that are in S. In terms of the probability wheel, the event $M_S = m_s$ means that precisely m_s future observations fall in a slice allocated to S. Based on the NPI model [7], there are

$$L = \sum_{j \in OS} (n_j - 1) + (2r + l - K)^+ \quad (8)$$

slices of the wheel which must be assigned to a category in S.

In this section, we will consider the general case where m may take any value. We will focus on the event that M_S reaches a certain threshold value, ie. 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. As before, we derive the most conservative bounds possible within the NPI framework.

First we consider the lower probability. 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. This involves counting all arrangements such that R observations fall in a slice 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, and 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 $\binom{L-1+R}{R}$ 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 (1) there are $\binom{n-L-1+(m-R)}{m-R}$ different arrangements of the $m-R$ observations within these slices.

Multiplying the above binomial coefficients tells us the minimum number of arrangements for which $M_S = R$. We can now sum over all relevant values of R, leading to the lower probability

$$\underline{P}(M_S \geq m_s) = \binom{n+m-1}{m}^{-1} \sum_{R=m_S}^{m} \binom{L-1+R}{R} \times \binom{n-L-1+(m-R)}{m-R}. \tag{9}$$

We assume $0 < L < n$, because $L = 0$ leads to lower probability zero. We also assume $m_s > 0$.

Now we consider the upper probability, which means 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 a slice which must be assigned to S. As explained above, there are a total of $\binom{L-1+R}{R}\binom{n-L-1+(m-R)}{m-R}$ arrangements such as this.

However, there are other arrangements which must now be included. We can now make use of the optional slices, ie. those slices which we can choose to assign either to S or to its complement. By considering the difference between the lower and upper probabilities given by the NPI model [7], we see that there are

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

optional slices. 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.

Suppose 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 one of the optional slices must be counted when calculating the 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 a certain Y, there are $\binom{L-1+Y}{Y}$ different arrangements of the Y observations within the slices which must be assigned to S. Also, there are $\binom{Q-1+T}{T}$ different arrangements of the T observations within the optional slices. Finally, there are $\binom{n-L-Q-1+(m-Y-T)}{m-Y-T}$ different arrangements of the other observations within the remaining slices of the wheel.

Combining these three binomial coefficents gives us the following upper probability:

$$\overline{P}(M_S \geq m_S) = \binom{n+m-1}{m}^{-1} [\sum_{R=m_S}^{m} \binom{L-1+R}{R} \times \binom{n-L-1+(m-R)}{m-R} + \sum_{Y=0}^{m_S-1} \sum_{T=m_S-Y}^{m-Y} \binom{L-1+Y}{Y}\binom{Q-1+T}{T} \times \binom{n-L-Q-1+(m-Y-T)}{m-Y-T}]. \tag{10}$$

As before, we assume $L > 0$ and $m_s > 0$. It is also 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 $\overline{P}(M_S \geq m_S) = 1$. In the case $L = 0$, the formula reduces to

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

Example 3.2. *Consider the data set in Example 2.2, where our possible categories are blue (B), red (R), yellow (Y) and green (G) and we have seen 5 observations including 2 B, 1 G, 1 Y and 1 R.*

We use inferences about three future observations, and we want to find the probability that at least one of these is in the subset $S = \{B, G\}$. To find the lower probability of this event, we use (9) with $m_S = 1$. We find that

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

and

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

for this example, and we also know that $n = 5$ and $m = 3$. Using these values we find that

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

When finding the upper probability, 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 want to find the probability that at least two of the three future observations are in S. We now apply (9) with $m_S = 2$, and we find that

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

As before, every slice on the wheel can be assigned to S, and 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 3.2. *For general m, when selecting an optimal subset of categories (see Introduction for our optimality criteria), categories should always be added to the subset in decreasing order of number of observations in the data set.*

Proof. Our aim is to select the subset which has the highest lower probability $\underline{P}(M_S \geq m_s)$ for some given value m_s. L is the only variable in this formula which 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 will consider two consecutive values of L. Consider the ratio

$$\frac{\underline{P}(M_S \geq m_s | L)}{\underline{P}(M_S \geq m_s | L + 1)}. \tag{11}$$

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 lower probability. If

$$\frac{\binom{L-1+R}{R}\binom{n-L-1+(m-R)}{m-R}}{\binom{L+R}{R}\binom{n-L+(m-R)}{m-R}} \tag{12}$$

is less than 1 for every possible value of R, then (11) must always be less than 1. Using the identities of the binomial coefficients, we can rewrite (12) 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, $(m-R)$ must always be positive regardless of the value of R since m is the maximum value of R, and $(n-L)$ must always be positive since L will always be less than 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 (8), the inclusion of an unobserved category in S will add 0 to the first term in L and 1 to the second term in L. The inclusion of an observed category in S will add $\frac{n_j - 1}{n}$ to the first term in L and 2 to the second term in L. So we see that it is always optimal to include observed categories in S before unobserved ones. Additionally, we see that the observed categories which will increase L by the greatest amount are those with the largest values of n_j. It is therefore always optimal to add categories to S in decreasing order of n_j. □

The following example illustrates how Theorem 3.2 can be implemented when selecting subsets.

Example 3.3. *Suppose that we have 8 possible categories, which we label A to H. We have made 100 observations. The table below shows how many of these observations were in each category.*

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

We want to investigate subsets of these 8 categories, and we will do this by making inferences about 2 future observations. There are two events of interest here: first, the event that at least one of the two future observations is in some subset S, and second, the event that both of the two 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 3.2, we know that the categories will be added in decreasing order of number of observations. The table below shows the composition of each of the subsets.

i	S_i	$P(M_{S_i} \geq 1)$	$P(M_{S_i} \geq 2)$
1	A	$[0.4206, 0.4505]$	$[0.0594, 0.0695]$
2	A,B	$[0.6727, 0.7166]$	$[0.1873, 0.2234]$
3	A-C	$[0.8376, 0.8822]$	$[0.3624, 0.4378]$
4	A-D	$[0.9196, 0.9543]$	$[0.5204, 0.6257]$
5	A-E	$[0.9697, 0.9846]$	$[0.6903, 0.7754]$
6	A-F	$[0.9945, 0.9980]$	$[0.8655, 0.9220]$
7	A-G	$[0.9998, 1.0000]$	$[0.9802, 1.0000]$
8	A-H	$[1.0000, 1.0000]$	$[1.0000, 1.0000]$

Using (9) and (10) with $m_S = 1$, we can find the lower and upper probabilities that at least one of the two future observations will be in S_i for $i = 1, ..., 8$. Similarly, we can use (9) and (10) with $m_S = 2$ to find the lower and upper probabilities that both of the two future observations will be in S_i for $i = 1, ..., 8$. The above table shows these probabilities.

Suppose that we want to select a subset of minimal size such that there is at least a 50% lower probability that one or more of the future observations will belong to a category in that subset. Looking at the above table of probabilities for the event $(M_{S_i} \geq 1)$, we see that the first row which satisfies $\underline{P}(M_{S_i} \geq 1) \geq 0.5$ is the row corresponding to $i = 2$. We therefore select the subset $S_2 = \{A, B\}$.

However, now suppose that we want to select the smallest possible subset of categories such that there is at least a 50% lower probability that both of the future observations will belong to a category in that subset. We will now need to select a larger subset in order to achieve the minimally required probability. Looking at the above table for the event $(M_{S_i} \geq 2)$, we see that the first row which satisfies $\underline{P}(M_{S_i} \geq 2) \geq 0.5$ is the row corresponding to $i = 4$. We therefore select the subset $S_4 = \{A, B, C, D\}$.

4 Concluding remarks

Coolen and Augustin [7] proved strong consistency properties for NPI, including F-probability in Weichselberger's theory of interval probability [10], but only for inferences involving a single future observation. For the case with multiple future observations, considered in this paper, these properties have not yet been proved, as we have thus far only derived the lower and upper probabilities of specific events. We would need to derive general formulae in order to investigate such properties. This is an interesting and important topic for future research. Further related research topics include other applications of NPI for multinomial data, where for example applications to classification are being investigated. Detailed comparisons of the NPI methods to more established alternatives may provide further insight into their practical value.

Acknowledgements

The authors thank the referee for helpful comments that led to the improvement of this paper.

References

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

[2] Bechhofer, R., Santner, T. and Goldsman, D. (1995) *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons.* Wiley.

[3] 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.

[4] Coolen, F.P.A. and Coolen-Schrijner, P. (2006) Nonparametric predictive subset selection for proportions *Statistics and Probability Letters*, **76**, 1675-1684.

[5] Coolen, F.P.A. and Coolen-Schrijner, P. (2007) Nonparametric predictive comparison of proportions *Journal of Statistical Planning and Inference*, **137**, 23-33.

[6] Coolen, F.P.A. and Augustin, T. (2005) Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model *ISIPTA '05*, 125-134.

[7] 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.

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

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

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