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

Frank Coolen

Department of Mathematical Sciences,
University of Durham

Thomas Augustin

Department of Statistics; SFB 386
University of Munich

1. Outline

- Aim: ∗ predictive inference
  ∗ based on n multinomial data (no ordering, e.g. colours)
  ∗ without prior knowledge
  ∗ categories defined upon observations
    ◦ c_j observed category j
    ◦ DN_i **d**escribed **n**ew category i
    ◦ UN **u**ndescribed **n**ew category

- The IDM (Walley (1996, JRSSB)) is the widely used standard model

- Here: A nonparametric alternative based on $\mathbb{A}_{(n)}$

More detailed: Available data:

- n_j observations in category c_j, $j = 1, \ldots, k$, $\sum_{j=1}^{k} n_j = n$ categories defined upon observation:
  $n_j \geq 1$, and $1 \leq k \leq n$
  (adding category with 0 observations to data does not affect our inferences)

- Defined New categories: DN_i, for $i = 1, \ldots, l$, with $l \geq 0$ - explicitly specified categories (not yet observed) - if only one, also denoted as DN

- Unobserved New category: UN - any observation not in a previously observed category, includes categories DN_i

2. The Imprecise Dirichlet model (IDM(M)) I

- Essentially a robust Bayesian model: Given a multinomial likelihood, consider all Dirichlet priors (conjugate prior)
- vacuous prior probabilities
- predictive probability to see colour c_i in the next trial

$$P(Y_{n+1} = c_i | \text{previous data}) = \left[\frac{n_i}{n+s}, \frac{n_i + s}{n+s} \right]$$

- IDMM (Walley & Bernard (1999, tech. rep.)) gives the same one-step predictive probabilities
- alternative view (Seidenfeld & Wasserman (1996, Disc. Walley)): ϵ-contaminated relative frequencies

The IDM(M) II

- Many powerful applications
- Survey: Bernard (2005, IJAR)
  At ISIPTA ´05 papers by: Abellan & Moral, Piatti & Zaffalon & Trojani, Silva & Campello de Souza, Strobl, Utkin & Augustin
- Walley's fundamental principle: RIP
  (representation invariance principle)

  'Posterior upper and lower probabilities assigned to an observable event A should not depend on the sample space in which A and the previous observations are represented'

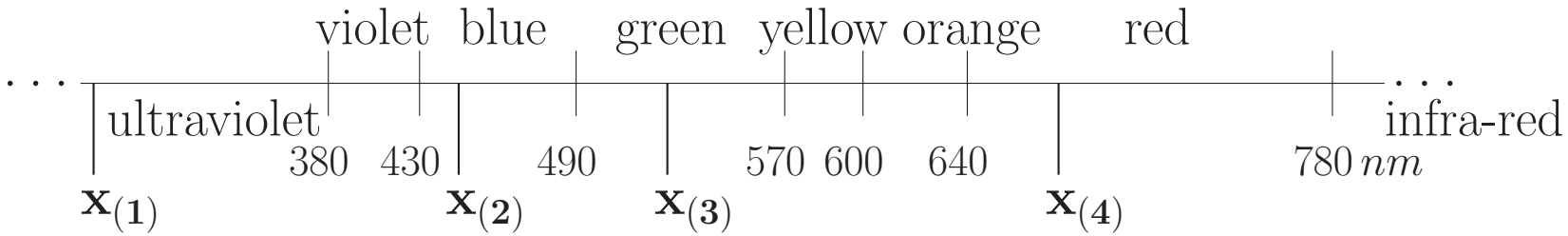
The IDM(M) III

- **Thesis**: The successful behaviour of the IDM is a result of the use of interval-valued instead of precise probability – and not of the specific model.

- Indeed, the IDM(M) has some strange and counter-intuitive features, raised by the discussants of Walley's (1996) paper, and even by Walley himself.

3. Predictive inference based on $\mathbb{A}_{(n)}$ – basic ideas

- *First step*: assume the colours to be ordered on the real line (e.g. wave length, sympathy index)
- observations (no ties):[1] $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$



- $A(n)$: for all $l = 0, \ldots, n$; post - data exchangeability:

$$P(Y_{n+1} \in (x_{(l)}, x_{(l+1)}) | \text{previous data}) = \frac{1}{n+1}$$

- Augustin & Coolen (2004), JSPI: totally monotone probability; many nice other properties

[1] $x_{(0)} := -\infty \quad x_{(n+1)} := \infty$

- Not of direct use for multinomial data: no ordering of the colours!
- *Second step*: consider a circle instead of the real line
  - probability wheel with coloured segments
  - $\Rightarrow \mathbb{A}_{(n)}$
  - "configuration" σ
  - similar properties of $P_\sigma(\cdot) = [\underline{P}_\sigma(\cdot), \overline{P}_\sigma(\cdot)]$: total monotonicity of $\underline{P}_\sigma(\cdot)$, F-probability with structure $\mathcal{M}_\sigma$
- The ordering is arbitrary $\Rightarrow$ consider *all possible* orderings and take the envelope over all configurations $\sigma \in \Sigma$:

  $$P(\cdot) = [\underline{P}(\cdot), \overline{P}(\cdot)] \quad \underline{P}(\cdot) = \min_{\sigma \in \Sigma} \underline{P}_\sigma(\cdot) \quad \overline{P}(\cdot) = \max_{\sigma \in \Sigma} \overline{P}_\sigma(\cdot)$$

- $P(\cdot)$ is F-probability, $\underline{P}(\cdot)$ is coherent.

The predictive probabilities of the general events

- Lower probability: minimum probability, for event of interest involving next observation, implied by this representation of data and $\mathcal{A}_{(n)}$

- Upper probability: similarly, the maximum probability that can be assigned to the event of interest, consistent with this representation of data and $\mathcal{A}_{(n)}$

Note: without the 'one category - one segment' assumption, our method would lead to vacuous lower and upper probabilities (i.e. 0 and 1, respectively)

$$\underline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \setminus \bigcup_{i=1}^{l} DN_i) =$$

$$\begin{cases} \dfrac{1}{n}\left(\displaystyle\sum_{s=1}^{r} n_{j_s} - r\right), \text{ for } k \geq 2r, \\ \dfrac{1}{n}\left(\displaystyle\sum_{s=1}^{r} n_{j_s} - r + \max(2r-k-l, 0)\right), \\ \qquad\qquad \text{for } r \leq k \leq 2r \end{cases}$$

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

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

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

4. Basic features in comparison with the IDM

- Asymptotics $n \longrightarrow \infty$: For any fixed sample space imprecision vanishes and the true proportions are learned correctly, too.
- Detailed examples $\longrightarrow$ poster
- Our inferences may depend on the data representation. Walley´s RIP is replaced by **two priniciples**:

1. Inferences based on different data representations must not be conflicting.

2. Parsimonity: Finer representations lead, ceteris paribus, to more imprecision.

1. Inferences based on different data representations must not be conflicting.

- For two representations R_1, R_2

$$P(Y_{n+1} \in A| \text{ data by } R_1) \cap P(Y_{n+1} \in A| \text{ data by } R_2) \neq \emptyset \ ,$$

  for all A describable by R_1 and R_2

- Note: in the case of classical probability this is equivalent to the RIP.

2. Parsimonity: Finer representations lead, ceteris paribus, to more imprecision.

- Let $\mathcal{A}_1$, $\mathcal{A}_2$ be the σ-fields produced by R_1 and R_2. If $\mathcal{A}_1 \supseteq \mathcal{A}_2$ then

$$P(A|\text{ data based on } R_1) \quad \supseteq \quad P(A|\text{ data based on } R_2)\,, \qquad \forall A \in \mathcal{A}_2$$

- A more detailed data representation allows more detailed inferences, but since this will imply less information on one or more categories, this may lead to more imprecision (certainly not less imprecision) for events expressable by all representations considered.

- **Seeing first outcome (f.o.) again**:
  - IDM: $\underline{P}(Y_2 = \text{ f.o. } | \text{ f.o. }) = \frac{1}{1+s} \quad (\frac{1}{2},\ s = 1,\ \frac{1}{3},\ s = 2)$
  - Here $\underline{P}(Y_2 = \text{ f.o. } | \text{ f.o. }) = 0$
  - our lower probability for this event is 0 - our lower probability for the next observation to belong to a category only becomes positive if that category has been observed at least twice.

- **Described and undescribed new categories**
  - $P(Y_{n+1} = DN) = \left[0; \frac{1}{n}\right]$

    but

    $P(Y_{n+1} = UN) = \left[0; \frac{k}{n}\right]$
  - Extreme cases:
    - If all n observations belong to the same category, then $P(Y_{n+1} = UN) = \left[0; \frac{1}{n}\right]$,
    - whereas if all n observations belong to different categories, we have $P(Y_{n+1} = UN) = [0; 1]$.
  - The lower and upper probabilities according to the IDM are $[0, \frac{s}{n+s}]$ for both these events, independent of other aspects of the data apart from n (RIP!).

Examples

Example 1: (Walley)
$n = 3$ observations: *1 Yellow, 1 Blue, 1 White*. Interest in '4th observation is Red or Yellow'

Consider 4 data representations:

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

RY denotes the category 'Red or Yellow', and O the category 'Other observed', that is here not 'Red or Yellow'.

Depending on the data representation, so on the definition of the observation categories, the event of interest is either denoted as $Y_4 = RY$ or as $Y_4 \in \{R, Y\}$.

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

Add 4th observation: *Red*

Now interested in Y_5 (using our model and $\mathcal{A}_{(4)}$)

(e) $D_e = (R : 1; Y : 1; B : 1; W : 1)$;
(f) $D_f = (RY : 2; O : 2)$;
(g) $D_g = (R : 1; Y : 1; O : 2)$;
(h) $D_h = (RY : 2; B : 1; W : 1)$.

These lead to the lower and upper probabilities:

(e) $P(Y_5 \in \{R, Y\}|D_e) = [0, 1]$;
(f) $P(Y_5 = RY|D_f) = [1/4, 3/4]$;
(g) $P(Y_5 \in \{R, Y\}|D_g) = [0, 3/4]$;
(h) $P(Y_5 = RY|D_h) = [1/4, 3/4]$.

Consider a fifth observation, in new category *Green*:

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

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

(*compare to (e)*) so with all observations in different categories, such upper probabilities are less than one if more than half of all observations belong to categories not in the event of interest.

Example 2: (Walley)

$n = 6$ observations: *1 Red, 3 Blue, 2 Green*

Event of interest: 7th observation is Red

We consider this, but also include the possibility that 7th observation belongs to a new category, and variations to data

$D_1 = (R : 1;\ B : 3;\ G : 2)$ leads to
$P(Y_7 = R|D_1) = [0, 2/6]$
and
$P(Y_7 \in \{R, UN\}|D_1) = [0, 3/6]$
$P(Y_7 \in \{R, DN\}|D_1) = [0, 3/6]$

Suppose that one observation was mistakenly classified as Blue, it should have been classified as Purple
$D_2 = (R : 1;\ B : 2;\ G : 2;\ P : 1)$ leads to

$P(Y_7 = R|D_2) = [0, 2/6]$
and
$\underline{P}(Y_7 \in \{R, UN\}|D_2) = \underline{P}(Y_7 \in \{R, DN\}|D_2) = 0$
which are the same as for D_1. However,
$\overline{P}(Y_7 \in \{R, UN\}|D_2) = 4/6$
and
$\overline{P}(Y_7 \in \{R, DN\}|D_2) = 3/6$
while for $l \geq 2$ we have
$\overline{P}(Y_7 \in \{R\} \cup \bigcup_{i=1}^{l} DN_i|D_2) = 4/6$

What if 2 Blue were distinguished: 1 Light Blue, 1 Dark Blue; and same for 2 Green:

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

$P(Y_7 = R|D_3) = [0, 2/6]$

and

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

are the same as for D_1 and D_2.

The upper probabilities for these latter two events are now

$\overline{P}(Y_7 \in \{R, UN\}|D_3) = 1$ and $\overline{P}(Y_7 \in \{R, DN\}|D_3) = \frac{3}{6}$

while

$\overline{P}(Y_7 \in \{R\} \cup \bigcup_{i=1}^{l} DN_i|D_3) = (2 + l)/6$ for $l = 2, 3$

and

$\overline{P}(Y_7 \in \{R\} \cup \bigcup_{i=1}^{l} DN_i|D_3) = 1$ for $l \geq 4$.

These upper probabilities correspond logically, by $\underline{P}(A) = 1 - \overline{P}(\bar{A})$, to the lower probabilities of the complementary

events, which is particularly clear for the event $Y_7 \in \{R, UN\}|D_3$, for which the complementary event has

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

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

$$P(Y_7 = UN|D_3) = [0, 1] \quad \text{and} \quad P(Y_7 = DN|D_3) = [0, 1/6]$$

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

5. More detailed technical results and their practical use

- $\underline{P}(\cdot)$ is two-monotone (but not totally monotone)

- The structure

$$\mathcal{M} = \left\{ p(\cdot) \,\middle|\, \underline{P}(\cdot) \leq p(\cdot) \leq \overline{P}(\cdot) \right\}$$

  of $P(\cdot)$ equals

$$\mathcal{M} = \text{conv} \left(\bigcup_{\sigma \in \Sigma} \mathcal{M}_\sigma \right)$$

- This means: All the information is exploited when working with $P(\cdot)$.
  Apply two-monotonicity $\Longrightarrow$ easy calculation of
  - intuitive conditional probabilities
  - lower and upper expectation / prevision by Choquet integration: For every $X : \Omega \to \mathbb{R}$

$$\underline{\mathbb{E}}_{\mathcal{M}} X = \sum_{A \subseteq \Omega} m(A) \min_{\omega \in A} X(\omega)$$

and

$$\overline{\mathbb{E}}_{\mathcal{M}} X = \sum_{A \subseteq \Omega} m(A) \max_{\omega \in A} X(\omega)$$

$\Longrightarrow$ direct application in decision making and classification.