

Chapter 2

Predictive Inference Under Exchangeability, and the Imprecise Dirichlet Multinomial Model

Gert de Cooman, Jasper De Bock and Márcio Diniz

Abstract Coherent reasoning under uncertainty can be represented in a very general manner by coherent sets of desirable gambles. In this framework, and for a given finite category set, coherent predictive inference under exchangeability can be represented using Bernstein coherent cones of multivariate polynomials on the simplex generated by this category set. This is a powerful generalisation of de Finetti's representation theorem allowing for both imprecision and indecision. We define an inference system as a map that associates a Bernstein coherent cone of polynomials with every finite category set. Many inference principles encountered in the literature can then be interpreted, and represented mathematically, as restrictions on such maps. We discuss two important inference principles: representation insensitivity—a strengthened version of Walley's representation invariance—and specificity. We show that there is a infinity of inference systems that satisfy these two principles, amongst which we discuss in particular the inference systems corresponding to (a modified version of) Walley and Bernard's imprecise Dirichlet multinomial models (IDMMs) and the Haldane inference system.

2.1 Introduction

This chapter deals with predictive inference for categorical variables. We are therefore concerned with a (possibly infinite) sequence of variables X_n that assume values in some finite set of categories A . After having observed a number \tilde{n} of them, and having found that, say $X_1 = x_1, X_2 = x_2, \dots, X_{\tilde{n}} = x_{\tilde{n}}$, we consider some subject's belief

G. de Cooman (✉) · J. De Bock · M. Diniz
Ghent University, SYSTeMS Research Group, Technologiepark–Zwijnaarde 914,
9052 Zwijnaarde, Belgium,
e-mail: gert.decooman@UGent.be

J. De Bock
e-mail: jasper.debock@UGent.be

M. Diniz
e-mail: marcio.diniz@UGent.be

model for the next \hat{n} variables $X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}$. In the probabilistic tradition—and we want to build on this tradition in the context of this chapter—this belief can be modelled by some conditional predictive probability mass function $p^{\hat{n}}(\cdot | x_1, \dots, x_{\check{n}})$ on the set $A^{\hat{n}}$ of possible values for these next variables. These probability mass functions can be used for prediction or estimation, for statistical inferences, and in decision making involving the uncertain values of these variables. In this sense, predictive inference lies at the heart of statistics, and of learning under uncertainty.

What connects these predictive probability mass functions for various values of \check{n} , \hat{n} and $(x_1, \dots, x_{\check{n}})$ are the requirements of *temporal consistency* and *coherence*. The former requires that when $n_1 \leq n_2$, $p^{n_1}(\cdot | x_1, \dots, x_{\check{n}})$ can be obtained from $p^{n_2}(\cdot | x_1, \dots, x_{\check{n}})$ through marginalisation; the latter essentially demands that these conditional probability mass functions should be connected with temporally consistent unconditional probability mass functions through Bayes's Rule.

A common assumption about the variables X_n is that they are *exchangeable*. De Finetti's famous representation theorem [4, 11] then states that the temporally consistent and coherent conditional and unconditional predictive probability mass functions associated with a countably infinite exchangeable sequence of variables in A are completely characterised by¹ a unique probability measure on the Borel sets of the simplex of all probability mass functions on A , called its *representation*.

This leads us to the central problem of predictive inference: since there is an infinity of such probability measures on the simplex, which one does a subject choose in a particular context, and how can a given choice be motivated and justified? The subjectivists of de Finetti's persuasion would answer that this question needs no answer: a subject's personal predictive probabilities are entirely his, and temporal consistency and coherence are the only requirements he should heed. Proponents of the logicist approach to predictive inference would try enunciating general inference principles in order to narrow down, and hopefully eliminate entirely, the possible choices for the representing probability measures on the simplex. Our point of view holds a compromise between the subjectivist and logicist positions: it should be possible for a subject to make assessments for certain predictive probabilities, and to combine these with certain inference principles he finds reasonable. Although this is not the topic of the present chapter, the inference systems we introduce in Sect. 2.6 provide an elegant framework and tools for making conservative predictive inferences that combine (local) subjective probability assessments with (general) inference principles.

This idea of *conservative probabilistic inference* brings us to a central idea in de Finetti's approach to probability [13]: a subject should be able to make certain probability assessments, and we can then consider these as bounds on so-called precise probability models. Calculating such most conservative but tightest bounds is indeed what de Finetti's fundamental theorem of prevision [13, 19] is about. The theory of imprecise probabilities [25, 28, 30] looks at conservative probabilistic inference precisely in this way: how can we calculate as efficiently as possible the

¹ ... unless the observed sequence has probability zero.

consequences—in the sense of most conservative tightest bounds—of making certain probability assessments. One advantage of imprecise probability models is that they allow for *imprecision*, or in other words, the use of *partial* probability assessments using bounding *inequalities* rather than equalities. In Sect. 2.2, we give a concise overview of the relevant ideas, models and techniques in the field of imprecise probabilities.

The present chapter, then, can be described as an application of ideas in imprecise probabilities to predictive inference. Its aim is to study—and develop a general framework for dealing with—coherent predictive inference using imprecise probability models. Using such models will also allow us to represent a subject’s indecision, which we believe is a natural state to be in when knowing, or having learned little, about the problem at hand. It seems important to us that theories of learning under uncertainty in general, and predictive inference in particular, start out with conservative, very imprecise and indecisive models when little has been learned, and become more precise and decisive as more observations come in.

Our work here builds on, but manages to reach much further than, an earlier chapter by one of the authors [9]. The main reason why it does so, is that we are now in a position to use a very powerful mathematical language to represent imprecise-probabilistic inferences: Walley’s [28] coherent sets of desirable gambles. Here, the primitive notions are not probabilities of events, nor expectations of random variables. The focus is rather on the question whether a gamble, or a risky transaction, is desirable to a subject—strictly preferred to the zero transaction, or status quo. And a basic belief model is now not a probability measure or lower prevision, but a *set of desirable gambles*.

Let us briefly summarise why, in the present chapter, we work with such sets as our basic uncertainty models for doing conservative probabilistic inference. Most importantly, and as we shall see in Sects. 2.2 and 2.3, marginalisation and conditioning are especially straightforward, and there are no issues whatsoever with conditioning on sets of (lower) probability zero. Furthermore, sets of desirable gambles provide an extremely expressive and general framework: It encompasses and subsumes as special cases both classical (or ‘precise’) probabilistic inference and inference in classical propositional logic [7].

So, now that we have argued why we want to use sets of desirable gambles to extend the existing probabilistic theory of predictive inference, let us explain in some detail how we intend to go about doing this. The basic building blocks are introduced in Sects. 2.2–2.8. As already indicated above, we give an overview of relevant notions and results concerning our imprecise probability model of choice—coherent sets of desirable gambles—in Sect. 2.2. In particular, we explain how to use them for conservative inference as well as conditioning; how to derive more commonly used models, such as lower previsions and lower probabilities, from them; and how they relate to precise probability models.

In Sect. 2.3, we explain how we can describe a subject’s beliefs about a sequence of variables in terms of predictive sets of desirable gambles, and the derived notion of predictive lower previsions. These imprecise probability models generalise the

above-mentioned predictive probability mass functions $p^{\hat{n}}(\cdot | x_1, \dots, x_{\hat{n}})$, and they constitute the basic tools we shall be working with. We also explain what are the proper formulations for the above-mentioned temporal consistency and coherence requirements in this more general context.

In Sect. 2.4, we discuss a number of inference principles that we believe could be reasonably imposed on predictive inferences, and we show how to represent them mathematically in terms of predictive sets of desirable gambles and lower previsions. *Representation insensitivity* means that predictive inferences remain essentially unchanged when we transform the set of categories, or in other words that they are essentially insensitive to the choice of representation—the category set. Another inference principle we look at imposes the so-called *specificity* property: when predictive inference is specific, then for a specific question involving a restricted number of categories, a more general model can be replaced by a more specific model that deals only with the categories of interest, and will produce the same relevant inferences [2].

The next important step is taken in Sect. 2.5, where we recall from the literature [8, 10] how to deal with exchangeability when our predictive inference models are imprecise. We recall that de Finetti’s representation theorem can be significantly generalised. In this case, the temporal consistent and coherent predictive sets of desirable gambles are completely characterised by a set of (multivariate) polynomials on the simplex of all probability mass functions on the category set. This set of polynomials must satisfy a number of properties, which taken together define the notion of *Bernstein coherence*. It serves completely the same purpose as the representing probability measure: it completely determines, and conveniently and densely summarises, all predictive inferences. This is the reason why the rest of the developments in the chapter are expressed in terms of such Bernstein coherent sets of polynomials.

We introduce coherent inference systems in Sect. 2.6 as maps that associate with any finite set of categories a Bernstein coherent set of polynomials on the simplex of probability mass functions on that set. The inference principles in Sect. 2.4 impose connections between predictive inferences for different category sets, so we can represent such inference principles mathematically as restrictions on coherent inference systems, which is the main topic of Sect. 2.7.

The material in Sects. 2.8–2.10 shows, by producing explicit examples, that there are quite a few different types—even uncountable infinities—of coherent inference systems that are both representation insensitive and specific. We discuss the vacuous inference system in Sect. 2.8, the family of IDMM inference systems in Sect. 2.9 and the Haldane inference system in Sect. 2.10.

In the Conclusion (Sect. 2.11) we point to a number of surprising consequences of our results, and discuss avenues for further research.

2.2 Imprecise Probability Models

In this section, we give a concise overview of imprecise probability models for representing, and making inferences and decisions under, uncertainty.

We shall focus on sets of desirable gambles as our uncertainty models of choice, because they are the most powerful, expressive and general models at hand, because they are very intuitive to work with—though unfortunately less familiar to most people not closely involved in the field—and very importantly, because they avoid problems with conditioning on sets of (lower) probability zero. For more details, we refer to Refs. [1, 5, 8, 21, 28]. We shall, of course, also briefly mention derived results in terms of the more familiar language of (lower) previsions and probabilities.

We consider a variable X that assumes values in some possibility space A . We model a subject's beliefs about the value of X by looking at which gambles on this variable the subject finds *desirable*, meaning that he strictly prefers them to the zero gamble—the status quo. This is a very general approach, that extends the usual rationalist and subjectivist approach to probabilistic modelling to allow for indecision and imprecision.

A *gamble* is a (bounded) real-valued function f on A . It is interpreted as an uncertain reward $f(X)$ that depends on the value of X , and is expressed in units of some predetermined linear utility. It represents the reward the subject gets in a transaction where first the actual value x of X is determined, and then the subject receives the amount of utility $f(x)$ —which may be negative, meaning he has to pay it. Throughout the chapter, we shall use the device of writing $f(X)$ when we want to make clear what variable the gamble f depends on. *Events* are subsets of the possibility space A . With any event $B \subseteq A$ we can associate a special gamble \mathbb{I}_B , called its *indicator*, which assumes the value 1 on B and 0 elsewhere.

We denote the set of all gambles on A by $\mathcal{G}(A)$. It is a linear space under point-wise addition of gambles, and point-wise multiplication of gambles with real numbers. For any subset \mathcal{A} of $\mathcal{G}(A)$, $\text{posi}(\mathcal{A})$ is the set of all positive linear combinations of gambles in \mathcal{A} : $\text{posi}(\mathcal{A}) := \{\sum_{k=1}^n \lambda_k f_k : f_k \in \mathcal{A}, \lambda_k \in \mathbb{R}_{>0}, n \in \mathbb{N}\}$. Here, \mathbb{N} is the set of natural numbers (without zero), and $\mathbb{R}_{>0}$ is the set of all positive real numbers. A *convex cone* of gambles is a subset \mathcal{A} of $\mathcal{G}(A)$ that is closed under positive linear combinations, meaning that $\text{posi}(\mathcal{A}) = \mathcal{A}$. For any two gambles f and g on A , we write ' $f \geq g$ ' if $(\forall x \in A) f(x) \geq g(x)$, and ' $f > g$ ' if $f \geq g$ and $f \neq g$. A gamble $f > 0$ is called *positive*. A gamble $g \leq 0$ is called *non-positive*. $\mathcal{G}_{>0}(A)$ denotes the convex cone of all positive gambles, and $\mathcal{G}_{\leq 0}(A)$ the convex cone of all non-positive gambles.

We collect the gambles that a subject finds desirable—strictly prefers to the zero gamble—into his *set of desirable gambles*, and we shall take such sets as our basic uncertainty models. Of course, they have to satisfy certain rationality criteria:

Definition 1 [Coherence] A set of desirable gambles $\mathcal{D} \subseteq \mathcal{G}(A)$ is called *coherent* if it satisfies the following requirements:

- D1. $0 \notin \mathcal{D}$;
- D2. $\mathcal{G}_{>0}(A) \subseteq \mathcal{D}$;
- D3. $\mathcal{D} = \text{posi}(\mathcal{D})$.

Requirement D3 turns \mathcal{D} into a *convex cone*. Due to D2, it includes $\mathcal{G}_{>0}(A)$; by D1–D3, it *avoids non-positivity*:

- D4. if $f \leq 0$ then $f \notin \text{posi}(\mathcal{D})$, or equivalently $\mathcal{G}_{\leq 0}(A) \cap \text{posi}(\mathcal{D}) = \emptyset$.

$\mathcal{G}_{>0}(A)$ is the smallest coherent subset of $\mathcal{G}(A)$. This so-called *vacuous model* therefore, reflects minimal commitments on the part of the subject: if he knows absolutely nothing about the likelihood of the different outcomes, he will only strictly prefer to zero those gambles that never decrease his wealth and have some possibility of increasing it.

Let us suppose that our subject has a coherent set \mathcal{D} of desirable gambles on A , expressing his beliefs about the value that a variable X assumes in A . We can then ask what his so-called *updated* set $\mathcal{D} \downarrow B$ of desirable gambles on B would be were he to receive the additional information—and nothing more—that X actually belongs to some subset B of A . The *updating*, or *conditioning*, *rule* for sets of desirable gambles states that:

$$g \in \mathcal{D} \downarrow B \Leftrightarrow g \mathbb{I}_B \in \mathcal{D} \text{ for all gambles } g \text{ on } B. \quad (2.1)$$

It states that the gamble g is desirable to a subject were he to observe that $X \in B$ if and only if the *called-off gamble* $g \mathbb{I}_B$ is desirable to him. This called-off gamble $g \mathbb{I}_B$ is the gamble on the variable X that gives a zero reward—is called off—unless $X \in B$, and in that case reduces to the gamble g on the new possibility space B . The updated set $\mathcal{D} \downarrow B$ is a set of desirable gambles on B that is still coherent, provided that \mathcal{D} is [8]. We refer to Refs. [5, 21, 22] for detailed discussions of updating sets of desirable gambles.

We now use coherent sets of desirable gambles to introduce derived concepts, such as coherent lower previsions and probabilities. Given a coherent set of desirable gambles \mathcal{D} , the functional \underline{P} defined on $\mathcal{G}(A)$ by

$$\underline{P}(f) := \sup\{\mu \in \mathbb{R} : f - \mu \in \mathcal{D}\} \text{ for all } f \in \mathcal{G}(A), \quad (2.2)$$

is a *coherent lower prevision* [25] [Theorem 3.8.1]. The conjugate upper prevision \overline{P} is defined by $\overline{P}(f) := \inf\{\mu \in \mathbb{R} : \mu - f \in \mathcal{D}\} = -\underline{P}(-f)$. For any gamble f , $\underline{P}(f)$ is called the *lower prevision* of f , and for any event B , $\underline{P}(\mathbb{I}_B)$ is also denoted by $\underline{P}(B)$, and called the *lower probability* of B . Similarly for upper previsions and upper probabilities.

The coherent conditional model $\mathcal{D} \downarrow B$, with B a non-empty subset of A , induces a *conditional lower prevision* $\underline{P}(\cdot | B)$ on $\mathcal{G}(B)$, by applying Eq. 2.2:

$$\begin{aligned} \underline{P}(g|B) &:= \sup\{\mu \in \mathbb{R} : g - \mu \in \mathcal{D} \downarrow B\} = \sup\{\mu \in \mathbb{R} : [g - \mu] \mathbb{I}_B \in \mathcal{D}\} \\ &\text{for all gambles } g \text{ on } B. \end{aligned} \quad (2.3)$$

It is not difficult to show [25] that \underline{P} and $\underline{P}(\cdot | B)$ are related through the following coherence condition:

$$\underline{P}([g - \underline{P}(g|B)]\mathbb{I}_B) = 0 \text{ for all } g \in \mathcal{G}(B), \quad (\text{GBR})$$

called the *Generalised Bayes Rule*. This rule allows us to infer $\underline{P}(\cdot | B)$ uniquely from \underline{P} , provided that $\underline{P}(B) > 0$. Otherwise, there are an infinity of coherent lower previsions $\underline{P}(\cdot | B)$ that are coherent with \underline{P} in the sense that they satisfy GBR.

Coherent sets of desirable gambles are more informative than coherent lower previsions: a gamble with positive lower prevision is always desirable and one with a negative lower prevision never, but a gamble with zero lower prevision lies on the border of the set of desirable gambles, and the lower prevision does not generally provide information about the desirability of such gambles. If such border behaviour is important—and it is when dealing with conditioning on events with zero (lower) probability [5, 21, 22, 28]—it is useful to work with sets of desirable gambles rather than lower previsions, because as Eqs. 2.1 and 2.3 tell us, they allow us to derive unique conditional models from unconditional ones.

When the lower and the upper prevision coincide on all gambles, then the real functional P defined on $\mathcal{G}(A)$ by $P(f) := \underline{P}(f) = \overline{P}(f)$ for all $f \in \mathcal{G}(A)$ is a *linear prevision*. In the particular case that A is finite, this means that it corresponds to the expectation operator associated with a probability mass function p : $P(f) = \sum_{x \in A} f(x)p(x) := E_p(f)$, where $p(x) := P(\mathbb{I}_{\{x\}})$ for all $x \in A$.

2.3 Predictive Inference

Predictive inference, in the specific sense we are focussing on here, considers a number of variables X_1, \dots, X_n assuming values in the same category set A —we define a *category set* as any non-empty *finite* set. We start our discussion of predictive inference models in the most general and representationally powerful language: coherent sets of desirable gambles, as introduced in the previous section.

Predictive inference assumes generally that a number \check{n} of observations have been made, so we know the values $\check{x} = (x_1, \dots, x_{\check{n}})$ of the first \check{n} variables $X_1, \dots, X_{\check{n}}$. Based on this *observation sample* \check{x} , a subject then has a posterior *predictive model* $\mathcal{D}_A^{\hat{n}} \rfloor \check{x}$ for the values that the next \hat{n} variables $X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}}$ assume in $A^{\hat{n}}$. $\mathcal{D}_A^{\hat{n}} \rfloor \check{x}$ is a coherent set of desirable gambles $f(X_{\check{n}+1}, \dots, X_{\check{n}+\hat{n}})$ on $A^{\hat{n}}$. Here, we assume that $\hat{n} \in \mathbb{N}$. On the other hand, we want to allow that $\check{n} \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, which is the set of all natural numbers with zero: we also want to be able to deal with the case where no previous observations have been made. In that case, we call the corresponding model $\mathcal{D}_A^{\hat{n}}$ a *prior predictive model*. Of course, technically speaking, $\check{n} + \hat{n} \leq n$.

As we said, the subject may also have a prior, unconditional model, for when no observations have yet been made. In its most general form, this will be a coherent set \mathcal{D}_A^n of desirable gambles $f(X_1, \dots, X_n)$ on A^n , for some $n \in \mathbb{N}$. Our subject may also have a coherent set $\mathcal{D}_A^{\hat{n}}$ of desirable gambles $f(X_1, \dots, X_n)$ on A^n , where

$\hat{n} \leq n$; and the sets $\mathcal{D}_A^{\hat{n}}$ and \mathcal{D}_A^n then be related to each other through the following *marginalisation*, or *temporal consistency*, requirement:

$$f(X_1, \dots, X_{\hat{n}}) \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f(X_1, \dots, X_n) \in \mathcal{D}_A^n \text{ for all gambles } f \text{ on } A^{\hat{n}}. \quad (2.4)$$

In this expression, and throughout this chapter, we identify a gamble f on $A^{\hat{n}}$ with its *cylindrical extension* f' on A^n , defined by $f'(x_1, \dots, x_{\hat{n}}, \dots, x_n) := f(x_1, \dots, x_{\hat{n}})$ for all $(x_1, \dots, x_n) \in A^n$. If we introduce the marginalisation operator $\text{marg}_{\hat{n}}(\cdot) := \cdot \cap \mathcal{G}(A^{\hat{n}})$, then the temporal consistency condition can also be rewritten simply as $\mathcal{D}_A^{\hat{n}} = \text{marg}_{\hat{n}}(\mathcal{D}_A^n) = \mathcal{D}_A^n \cap \mathcal{G}(A^{\hat{n}})$.

Prior (unconditional) predictive models \mathcal{D}_A^n and posterior (conditional) ones $\mathcal{D}_A^{\hat{n}} \downarrow \tilde{\mathbf{x}}$ must also be related through the following *updating* requirement:

$$f(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}}) \in \mathcal{D}_A^{\hat{n}} \downarrow \tilde{\mathbf{x}} \Leftrightarrow f(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}}) \mathbb{I}_{\{\tilde{\mathbf{x}}\}}(X_1, \dots, X_{\tilde{n}}) \in \mathcal{D}_A^n \\ \text{for all gambles } f \text{ on } A^{\hat{n}}, \quad (2.5)$$

which is a special case of Eq. 2.1: the gamble $f(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}})$ is desirable after observing a sample $\tilde{\mathbf{x}}$ if and only if the gamble $f(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}}) \mathbb{I}_{\{\tilde{\mathbf{x}}\}}(X_1, \dots, X_{\tilde{n}})$ is desirable before any observations are made. This called-off gamble is the gamble that gives zero reward—is called off—unless the first \tilde{n} observations are $\tilde{\mathbf{x}}$, and in that case reduces to the gamble $f(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}})$ on the variables $X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}}$. The updating requirement is a generalisation of Bayes's Rule for updating, and in fact reduces to it when the sets of desirable gambles lead to (precise) probability mass functions [7, 28]. But contrary to Bayes's Rule for probability mass functions, the updating rule 2.5 for coherent sets of desirable gambles clearly does not suffer from problems when the conditioning event has (lower) probability zero: it allows us to infer a unique conditional model from an unconditional one, regardless of the (lower or upper) probability of the conditioning event.

As explained in Sect. 2.2, we can use the relationship 2.2 to derive *prior* (unconditional) *predictive lower previsions* $\underline{P}_A^{\hat{n}}(\cdot)$ on $\mathcal{G}(A^{\hat{n}})$ from the prior sets $\mathcal{D}_A^{\hat{n}}$ through:

$$\underline{P}_A^{\hat{n}}(f) := \sup\{\mu \in \mathbb{R} : f - \mu \in \mathcal{D}_A^{\hat{n}}\} \text{ for all gambles } f \text{ on } A^{\hat{n}},$$

and *posterior* (conditional) *predictive lower previsions* $\underline{P}_A^{\hat{n}}(\cdot \downarrow \tilde{\mathbf{x}})$ on $\mathcal{G}(A^{\hat{n}})$ from the posterior sets $\mathcal{D}_A^{\hat{n}} \downarrow \tilde{\mathbf{x}}$ through:

$$\underline{P}_A^{\hat{n}}(f \downarrow \tilde{\mathbf{x}}) := \sup\{\mu \in \mathbb{R} : f - \mu \in \mathcal{D}_A^{\hat{n}} \downarrow \tilde{\mathbf{x}}\} \text{ for all gambles } f \text{ on } A^{\hat{n}}.$$

We also want to condition predictive lower previsions on the additional information that $(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}}) \in B^{\hat{n}}$, where B is some proper subset of A . Using the ideas in Sects. 2.2, this leads for instance to the following lower prevision:

$$\underline{P}_A^{\hat{n}}(g \downarrow \tilde{\mathbf{x}}, B^{\hat{n}}) := \sup\{\mu \in \mathbb{R} : [g - \mu] \mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}} \downarrow \tilde{\mathbf{x}}\} \text{ for all gambles } g \text{ on } B^{\hat{n}}, \quad (2.6)$$

which is the lower prevision $\underline{P}_A^{\hat{n}}(\cdot \downarrow \tilde{\mathbf{x}})$ conditioned on the event $B^{\hat{n}}$.

2.4 Principles for Predictive Inference

So far, we have introduced coherence, marginalisation and updating as basic requirements of rationality that prior and posterior predictive inference models must satisfy. In addition to these, we now also consider a number of further conditions, which have been suggested by a number of authors as reasonable properties—or requirements—for predictive inference models.

We shall call *representation insensitivity* the combination of pooling, renaming and category permutation invariance; see Ref. [9] for more information. It means that predictive inferences remain essentially unchanged when we transform the set of categories, or in other words that they are essentially insensitive to the choice of representation—the category set. It is not difficult to see that representation insensitivity can be formally characterised as follows. Consider two category sets A and B such that there is a so-called *relabelling map* $\rho : A \rightarrow B$ that is *onto*, such that $B = \rho(A) := \{\rho(x) : x \in A\}$. Then with a sample \mathbf{x} in A^n , there corresponds a transformed sample $\rho\mathbf{x} := (\rho(x_1), \dots, \rho(x_n))$ in B^n . And with any gamble f on B^n there corresponds a gamble $f \circ \rho$ on A^n .

Representation insensitivity: For all category sets A and B such that there is an onto map $\rho : A \rightarrow B$, all $\tilde{n}, \hat{n} \in \mathbb{N}$ considered, all $\tilde{\mathbf{x}} \in A^{\tilde{n}}$ and all gambles f on $B^{\hat{n}}$:

$$\underline{P}_A^{\hat{n}}(f \circ \rho) = \underline{P}_B^{\hat{n}}(f) \text{ and } \hat{n}_A(f \circ \rho)\tilde{\mathbf{x}} = \underline{P}_B^{\hat{n}}(f|\rho\tilde{\mathbf{x}}), \quad (\text{RI1})$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f \circ \rho \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \text{ and } f \circ \rho \in \mathcal{D}_A^{\hat{n}} \downarrow \tilde{\mathbf{x}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \downarrow \rho\tilde{\mathbf{x}}. \quad (\text{RI2})$$

There is another peculiar, but in our view intuitively appealing, potential property of predictive inferences. Assume that in addition to observing a sample of observations $\tilde{\mathbf{x}}$ of \tilde{n} observations in a category set A , our subject comes to know or determine in some way that the \hat{n} following observations will belong to a proper subset B of A , and nothing else—we might suppose for instance that an observation of $(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}})$ has been made, but that it is imperfect, and only allows him to conclude that $(X_{\tilde{n}+1}, \dots, X_{\tilde{n}+\hat{n}}) \in B^{\hat{n}}$.

We can then make the following requirement, which uses models conditioned on the event $B^{\hat{n}}$, as introduced through Eqs. 2.1, 2.3 and 2.6.

Specificity: For all category sets A and B such that $B \subseteq A$, all $\tilde{n}, \hat{n} \in \mathbb{N}$ considered, all $\tilde{\mathbf{x}} \in A^{\tilde{n}}$ and all gambles f on $B^{\hat{n}}$:

$$\underline{P}_A^{\hat{n}}(f|B^{\hat{n}}) = \underline{P}_B^{\hat{n}}(f) \text{ and } \hat{n}_A(f|\tilde{\mathbf{x}}, B^{\hat{n}}) = \underline{P}_B^{\hat{n}}(f|\tilde{\mathbf{x}}\downarrow_B), \quad (\text{SP1})$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f \mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \text{ and } f \mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}} \downarrow \tilde{\mathbf{x}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \downarrow \tilde{\mathbf{x}}\downarrow_B, \quad (\text{SP2})$$

where $\tilde{\mathbf{x}}\downarrow_B$ is the tuple of observations obtained by eliminating from the tuple $\tilde{\mathbf{x}}$ all observations not in B . In these expressions, when $\tilde{\mathbf{x}}\downarrow_B$ is the empty tuple, so when

no observations in \check{x} are in B , the ‘posterior’ predictive model is simply taken to reduce to the ‘prior’ predictive model. Specificity [2, 3, 24] means that *the predictive inferences that a subject makes are the same as the ones he would get by focussing on the category set B , and at the same time discarding all the previous observations producing values outside B , in effect only retaining the observations that were inside B !* It is as if knowing that the future observations belong to B allows our subject to ignore all the previous observations that happened to lie outside B .

2.5 Adding Exchangeability to the Picture

We are now, for the remainder of this chapter, going to add two additional assumptions. The *first assumption* is that we are dealing with a *countably infinite sequence* of variables X_1, \dots, X_n, \dots that assume values in the same category set A . For our predictive inference models, this means that there is a sequence \mathcal{D}_A^n of coherent sets of desirable gambles on A^n , $n \in \mathbb{N}$. The *second assumption* is that this sequence of variables is *exchangeable*, which means, roughly speaking, that the subject believes that the order in which these variables are observed, or present themselves, has no influence on the decisions and inferences he will make regarding these variables.

In this section, we explain succinctly how to deal with these assumptions technically, and what their consequences are for the predictive models we are interested in. For a detailed discussion and derivation of the results presented here, we refer to Refs. [8, 10].

We begin with some useful notation, which will be employed numerous times in what follows. Consider any element $\alpha \in \mathbb{R}^A$. We consider α as an A -tuple, with as many (real) components $\alpha_x \in \mathbb{R}$ as there are categories x in A . For any subset $B \subseteq A$, we then denote by $\alpha_B := \sum_{x \in B} \alpha_x$ the sum of its components over B .

Consider an arbitrary $n \in \mathbb{N}$. We denote by $\mathbf{x} = (x_1, \dots, x_n)$ a generic, arbitrary element of A^n . \mathcal{P}^n is the set of all permutations π of the index set $\{1, \dots, n\}$. With any such permutation π , we can associate a permutation of A^n , also denoted by π , and defined by $(\pi\mathbf{x})_k := x_{\pi(k)}$, or in other words, $\pi(x_1, \dots, x_n) := (x_{\pi(1)}, \dots, x_{\pi(n)})$. Similarly, we lift π to a permutation π^t of $\mathcal{G}(A^n)$ by letting $\pi^t f := f \circ \pi$, so $(\pi^t f)(\mathbf{x}) := f(\pi\mathbf{x})$. The permutation invariant atoms $[\mathbf{x}] := \{\pi\mathbf{x} : \pi \in \mathcal{P}^n\}$, $\mathbf{x} \in A^n$ are the smallest permutation invariant subsets of A^n .

We now introduce the *counting map* $\mathbf{T} : A^n \rightarrow \mathcal{N}_A^n : \mathbf{x} \mapsto \mathbf{T}(\mathbf{x})$, where the *count vector* $\mathbf{T}(\mathbf{x})$ is the A -tuple with components $T_z(\mathbf{x}) := |\{k \in \{1, \dots, n\} : x_k = z\}|$ for all $z \in A$, and the set of possible *count vectors* for n observations in A is given by $\mathcal{N}_A^n := \{\mathbf{m} \in \mathbb{N}_0^A : \mathbf{m}_A = n\}$. So, $T_z(\mathbf{x})$ is the number of times the category z appears in the sample \mathbf{x} . If $\mathbf{m} = \mathbf{T}(\mathbf{x})$, then $[\mathbf{x}] = \{\mathbf{y} \in A^n : \mathbf{T}(\mathbf{y}) = \mathbf{m}\}$, so the atom $[\mathbf{x}]$ is completely determined by the single count vector \mathbf{m} of all its elements, and is therefore also denoted by $[\mathbf{m}]$.

Interdisciplinary Bayesian Statistics

EBEB 2014

Polpo de Campos, A.; Neto, F.L.; Ramos Rifo, L.; Stern,

J.M.; Lauretto, M. (Eds.)

2015, XVIII, 366 p. 67 illus., 45 illus. in color., Hardcover

ISBN: 978-3-319-12453-7