Immediate prediction under exchangeability and representation insensitivity

Gert de Cooman
Ghent University
SYSTeMS Research Group
gert.decooman@ugent.be

Enrique Miranda
Rey Juan Carlos University
Dept. of Statistics and O.R.
enrique.miranda@urjc.es

Erik Quaeghebeur
Ghent University
SYSTeMS Research Group
erik.quaeghebeur@ugent.be

Abstract

We consider immediate predictive inference, where a subject, using a number of observations of a finite number of exchangeable random variables, is asked to coherently model his beliefs about the next observation, in terms of a predictive lower prevision. We study when such predictive lower previsions are representation insensitive, meaning that they are essentially independent of the choice of the (finite) set of possible values for the random variables. Such representation insensitive predictive models have very interesting properties, and among such models, the ones produced by the Imprecise Dirichlet-Multinomial Model are quite special in a number of ways.

Keywords. Predictive inference, immediate prediction, lower prevision, coherence, exchangeability, representation invariance, representation insensitivity, Imprecise Dirichlet-Multinomial Model, Johnson's sufficientness postulate.

1 Introduction

Consider a subject who is making $N > 0$ successive observations of a certain phenomenon. We represent these observations by N random variables $X_1, \dots, X_N$. By *random variable*, we mean a variable about whose value the subject may entertain certain beliefs. We assume that at each successive instant k, the actual value of the random variables X_k can be determined in principle. To fix ideas, our subject might be drawing balls without replacement from an urn, in which case X_k could designate the colour of the k-th ball taken from the urn.

In the type of predictive inference we consider here, our subject in some way uses zero or more observations $X_1, \dots, X_n$ made previously, i.e., those up to a certain instant $n \in \{0, 1, \dots, N-1\}$, to predict, or make inferences about, the values of the future, or as yet unmade, observations $X_{n+1}, \dots, X_N$. Here, we only consider the problem of *immediate prediction*: he is only trying to predict, or make inferences about, the value of the next observation X_{n+1}.

We are particularly interested in the problem of making such predictive inferences under prior ignorance: initially, *before making any observation, our subject knows very little or nothing about what produces these observations*. In the urn example, this is the situation where he doesn't know the composition of the urn, e.g., how many balls there are, or what their colours are. What we do assume, however, is that our subject makes an assessment of *exchangeability* to the effect that the order in which a sequence of observations has been made does not matter for his predictions.

What a subject usually does, in such a situation, is to determine, beforehand, a (finite and non-empty) set $\mathscr{X}$ of possible values, also called *categories*, for the random variables X_k. It is then sometimes held, especially by advocates of a logical interpretation for probability, that our subject's beliefs should be represented by some given family of predictive probability mass functions. Such a predictive family is made up of real-valued maps $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on $\mathscr{X}$, which give, for each $n = 0, \dots, N-1$ and each $\boldsymbol{x} = (x_1, \dots, x_n)$ in $\mathscr{X}^n$, the (so-called *predictive*) probability mass function for the $(n+1)$-th observation, given the values $(X_1, \dots, X_n) = (x_1, \dots, x_n) = \boldsymbol{x}$ of the n previous observations. Any such family should in particular reflect the above-mentioned exchangeability assessment. Cases in point are the Laplace–Bayes Rule of Succession in the case of two categories [10], or Carnap's more general λ-calculus [2].

The inferences in Carnap's λ-calculus, to give but one example, can strongly depend on the number of elements in the set $\mathscr{X}$. This may well be considered undesirable. If for instance, we consider drawing balls from an urn, predictive inferences about whether the next ball will be '*red or green*' ideally should not depend on whether we assume beforehand that the possible categories are '*red*', '*green*', '*blue*' and '*any other colour*', or whether we take them to be '*red or green*', '*blue*', '*yellow*' and '*any other colour*'. This desirable property was called *representation invariance* by Peter Walley [14], who argued that it cannot be satisfied by a *precise* probability model, i.e., by a system consisting of a family of predictive probability mass functions $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for every $\mathscr{X}$, but that it is satisfied by the so-called Imprecise Dirichlet-Multinomial Model (or IDMM for short [15]). The IDMM can be seen as a special system of predictive *lower previsions* and it is a (predictive) cousin of the parametric Imprecise Dirichlet Model (or IDM [14]). Lower previsions are behavioural belief models that generalise the more classical Bayesian ones, such as probability mass functions, or previsions. We assume that the reader is familiar with at least the basic aspects of the theory of coherent lower previsions [13].

Here, we intend to study general systems of such predictive lower previsions. In Section 2, we give a general definition of such predictive systems and study a number of properties they can satisfy, such as coherence and exchangeability. In Section 3, we study the property of representation insensitivity for predictive systems, which is a stronger version of Walley's representation invariance, tailored to making inferences under prior ignorance. We show in Section 4 that there are representation insensitive and exchangeable predictive systems, by giving two examples. These two can be used to generate the mixing predictive systems, studied in Section 5. Among these, the ones corresponding to an IDMM take a special place, as they are the only ones to satisfy all the above-mentioned properties and an extra *specificity* property, related to behaviour under conditioning. In the Conclusions (Section 6), we list a number of interesting, but as of yet unresolved, questions.

2 Predictive families and systems

2.1 Families of predictive lower previsions

First assume that, before the subject starts making the observations X_k, he fixes a non-empty and finite set $\mathscr{X}$ of possible values for all the random variables X_k. Now suppose that he has observed the sequence of values $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathscr{X}^n$ of the first n random variables, or in other words, he knows that $X_k = x_k$ for $k = 1, \dots, n$. We want to represent his beliefs about the value of the next observation X_{n+1}, and the model we propose for this is a lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on the set $\mathscr{L}(\mathscr{X})$ of all gambles on $\mathscr{X}$. Let us first make clear what this means (see Walley's book [13] for more information).

A *gamble* f on $\mathscr{X}$ is a real-valued map on $\mathscr{X}$. It represents an uncertain reward, expressed in terms of some predetermined linear utility scale. So a gamble f yields a (possibly negative) reward of $f(x)$ utiles if the value of the next variable X_{n+1} turns out to be the category x in $\mathscr{X}$. The set of all gambles on $\mathscr{X}$ is denoted by $\mathscr{L}(\mathscr{X})$. The *lower prevision* $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ of any gamble f on $\mathscr{X}$ is the subject's supremum acceptable price for buying this gamble, or in other words, the highest s such that he accepts the uncertain reward $f(X_{n+1}) - p$ for all $p < s$, conditional on his having observed the values $\boldsymbol{x} = (x_1, \dots, x_n)$ for the first n variables $(X_1, \dots, X_n)$. His corresponding *predictive upper prevision*, or infimum selling price for f, is then given by the conjugacy relationship: $\overline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) = -\underline{P}^{n+1}_{\mathscr{X}}(-f|\boldsymbol{x})$.

A specific class of gambles is related to *events*, i.e., subsets A of $\mathscr{X}$. This is the class of indicators I_A that map any element of A to one and all other elements of $\mathscr{X}$ to zero. We identify events A with their indicators I_A. A lower prevision that is defined on (indicators of) events only is called a *lower probability*, and we often write $\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{x})$ instead of $\underline{P}^{n+1}_{\mathscr{X}}(I_A|\boldsymbol{x})$.

The *predictive lower prevision* $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$, which models beliefs about the value of the random variable X_{n+1} given the observations $(X_1, \dots, X_n) = \boldsymbol{x}$, is the real-valued functional on $\mathscr{L}(\mathscr{X})$ that assigns to any gamble f its predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$. We assume that the subject has such a predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for all $\boldsymbol{x}$ in $\mathscr{X}^n$ and all $n \in \{0, \dots, N-1\}$, where $N > 0$ is some fixed positive integer, representing the total number of observations we are interested in. For $n = 0$, there is some slight abuse of notation here, because we then actually have an unconditional predictive lower prevision $\underline{P}^1_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X})$ for the first observation X_1, and no observations have yet been made.

Definition 1 (Family of predictive lower previsions)**.** *Consider a finite and non-empty set of categories $\mathscr{X}$. An $\mathscr{X}$-family of predictive lower previsions, or* predictive $\mathscr{X}$-family *for short, for up to $N > 0$ observations is a set of predictive lower previsions*

$$\sigma^N_{\mathscr{X}} := \left\{\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) \colon \boldsymbol{x} \in \mathscr{X}^n \text{ and } n = 0, \dots, N-1\right\}.$$

It is useful to consider the special case, quite common in the literature, of a family of predictive lower previsions of which all members $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ are actually *linear previsions* $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$. This means that for each $n = 0, \dots, N-1$ and $\boldsymbol{x}$ in $\mathscr{X}^n$ there is some predictive *(probability) mass function* $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on $\mathscr{X}$ such that $\sum_{z \in \mathscr{X}} p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}) = 1$, for all z in $\mathscr{X}$, $p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}) \geq 0$, and for all gambles f on $\mathscr{X}$

$$P^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) = \sum_{z \in \mathscr{X}} f(z) p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}).$$

Such linear previsions are the Bayesian belief models usually encountered in the literature (see for instance de Finetti's book [7]). We can use Bayes's rule to combine these predictive mass functions into unique *joint mass functions* $p^n_{\mathscr{X}}$ on $\mathscr{X}^n := \times_{i=1}^n \mathscr{X}$, given by

$$p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(x_1, \dots, x_n) = \prod_{k=0}^{n-1} p^{k+1}_{\mathscr{X}}(x_{k+1}|x_1, \dots, x_k),$$

for all $\boldsymbol{x} = (x_1, \dots, x_n)$ in $\mathscr{X}^n$ and all $n = 1, \dots, N$. This also results in unique corresponding linear previsions (expectation operators) $P^n_{\mathscr{X}}$ defined for all f in $\mathscr{L}(\mathscr{X}^n)$ by

$$P^n_{\mathscr{X}}(f) = \sum_{\boldsymbol{x} \in \mathscr{X}^n} f(\boldsymbol{x}) p^n_{\mathscr{X}}(\boldsymbol{x}). \tag{1}$$

For $n = N$, we call $P^N_{\mathscr{X}}$ the *joint linear prevision* associated with the given predictive family of linear previsions. It models beliefs about the values that the random variables $(X_1,\dots,X_N)$ assume *jointly* in $\mathscr{X}^N$.

2.2 Systems of predictive lower previsions

When a subject is using a family of predictive lower previsions $\sigma^N_{\mathscr{X}}$, this means he has assumed beforehand that the random variables X_1, ..., X_N all take values in the set $\mathscr{X}$. It cannot, therefore, be excluded at this point that his inferences, as represented by the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$, strongly depend on the choice of the set of possible values $\mathscr{X}$. Any initial choice of $\mathscr{X}$ may lead to an essentially very different predictive family $\sigma^N_{\mathscr{X}}$. In order to be able to deal with this possible dependence mathematically, we now define predictive systems as follows.

Definition 2 (System of predictive lower previsions)**.** *Fix $N > 0$. If we consider for any finite and non-empty set of categories $\mathscr{X}$ a corresponding $\mathscr{X}$-family $\sigma^N_{\mathscr{X}}$ of predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$, we get a new collection*

$$\sigma^N := \left\{\sigma^N_{\mathscr{X}} \colon \mathscr{X} \text{ is a finite and non-empty set}\right\},$$

called a system of predictive lower previsions*, or* predictive system *for short, for up to N observations. We denote the set of all predictive systems for a given (fixed) N by Σ^N.*

It is such predictive systems that we are interested in, and whose properties we intend to study. Consider the set Σ^N of all predictive systems for up to N observations. For two such predictive systems σ^N and λ^N we say that σ^N is *less committal*, or *more conservative*, than λ^N, and we denote this by $\sigma^N \preceq \lambda^N$, if each predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in σ^N is *point-wise dominated* by the corresponding predictive lower prevision $\underline{Q}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in λ^N:

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) \le \underline{Q}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$$

for all gambles f on $\mathscr{X}$. The reason for this terminology should be clear: a subject using a predictive system λ^N will then be buying gambles f on $\mathscr{X}$ at supremum prices $\underline{Q}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ that are at least as high as the supremum prices $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ of a subject using predictive system σ^N.

The binary relation $\preceq$ on Σ^N is a partial order. A non-empty subset $\left\{\sigma^N_\gamma \colon \gamma \in \Gamma\right\}$ of Σ^N (where Γ is some index set) may have an infimum with respect to this partial order, and whenever it exists, this infimum corresponds to taking *lower envelopes*: if we fix $\mathscr{X}$, n and $\boldsymbol{x}$, then the corresponding predictive lower prevision in the infimum predictive system is the lower envelope $\inf_{\gamma\in\Gamma}\underline{P}^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$ of the corresponding predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$ in the predictive systems σ^N_γ, $\gamma \in \Gamma$.

2.3 Coherence requirements

We impose some consistency, or rationality, requirements on the members $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ of a system σ^N of predictive lower previsions.

Definition 3 (Coherence)**.** *A system of predictive lower previsions is called* coherent *if it is the infimum (or lower envelope) of a collection of systems of predictive linear previsions.*

This condition is equivalent to requiring, for each choice of $\mathscr{X}$, that the conditional lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for $n = 0,\dots,N-1$ and $\boldsymbol{x} \in \mathscr{X}^n$ should satisfy Walley's (joint) coherence condition.[1] This condition is in the present context also equivalent [12] to requiring that the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ by themselves should be *(separately) coherent*, meaning that for each finite and non-empty set $\mathscr{X}$, $n = 0,\dots,N-1$ and $\boldsymbol{x}$ in $\mathscr{X}^n$, $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ should satisfy

(C1) $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) \ge \inf f$;

(C2) $\underline{P}^{n+1}_{\mathscr{X}}(f+g|\boldsymbol{x}) \ge \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) + \underline{P}^{n+1}_{\mathscr{X}}(g|\boldsymbol{x})$;

(C3) $\underline{P}^{n+1}_{\mathscr{X}}(\lambda f|\boldsymbol{x}) = \lambda \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$;

for all gambles f and g on $\mathscr{X}$ and all real $\lambda \ge 0$.

2.4 Exchangeability and regular exchangeability

Next, we show how to formulate an assessment of *exchangeability* of the random variables X_1, ..., X_N in terms of a system of predictive lower previsions. A subject would make such an assessment if he believed that the order in which these variables are observed is not important. Let us make this idea more precise.

We begin with the definition of exchangeability for a *precise* predictive system, i.e., a system of predictive linear previsions. For each choice of $\mathscr{X}$, the precise $\mathscr{X}$-family $\sigma^N_{\mathscr{X}}$ has a unique joint linear prevision $P^N_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X}^N)$, defined by Equation (1). *We then call the precise predictive system exchangeable if all the associated joint linear previsions $P^N_{\mathscr{X}}$ are.* Formally [5, 7], consider the set of all permutations of $\{1,\dots,N\}$. With any such permutation π we can associate a permutation of $\mathscr{X}^N$, also denoted by π, that maps any $\boldsymbol{x} = (x_1,\dots,x_N)$ in $\mathscr{X}^N$ to $\pi\boldsymbol{x} := (x_{\pi(1)},\dots,x_{\pi(N)})$. Similarly, with any gamble f on $\mathscr{X}^N$, we can consider the permuted gamble $\pi f := f \circ \pi$, or in other words $(\pi f)(\boldsymbol{x}) = f(\pi\boldsymbol{x})$. We then require that $P^N_{\mathscr{X}}(\pi f) = P^N_{\mathscr{X}}(f)$ for any such permutation π and any gamble f on $\mathscr{X}^N$. Equivalently, in terms of the joint mass

[1] See Chapters 6 and 7, and also Section K3 (Williams's Theorem) in Walley's book [13]. Since the random variables X_k are assumed to only take on a finite number of values, Walley's coherence condition coincides with the one first suggested by Williams [16].

function $p^N_{\mathscr{X}}$, we require that $p^N_{\mathscr{X}}(\pi\boldsymbol{x}) = p^N_{\mathscr{X}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ in $\mathscr{X}^N$ and all permutations π.

We adopt the following definition of exchangeability for general predictive systems.

Definition 4 (Exchangeability)**.** *A system of predictive lower previsions is called* exchangeable *if it is the infimum (or lower envelope) of a collection of exchangeable systems of predictive linear previsions. We denote by $\langle \Sigma^N_{\mathrm{e}}, \preceq\rangle$ the set of all exchangeable predictive systems for up to N observations, with the same order relation $\preceq$ that we defined on $\langle \Sigma^N, \preceq\rangle$.*

The infimum (lower envelope) of any non-empty collection of exchangeable predictive systems is still exchangeable. This means that the partially ordered set $\langle \Sigma^N_{\mathrm{e}}, \preceq\rangle$ is a complete semi-lattice [3, Sections 3.19–3.20]. For reasons of mathematical convenience, we also introduce a stronger requirement.

Definition 5 (Regular exchangeability)**.** *A system of predictive lower previsions is called* regularly exchangeable *if it is the infimum (or lower envelope) of some collection σ^N_γ, $\gamma \in \Gamma$ of exchangeable systems of predictive linear previsions, such that for all finite and non-empty $\mathscr{X}$, all $\boldsymbol{x}$ in $\mathscr{X}^{N-1}$, and all γ in Γ,*

$$\begin{aligned} p^{N-1}_{\mathscr{X},\gamma}(\boldsymbol{x}) :=& P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}) \\ =& \prod_{k=0}^{N-2} p^{k+1}_{\mathscr{X},\gamma}(x_{k+1}|x_1,\dots,x_k) > 0. \end{aligned}$$

Of course, all regularly exchangeable predictive systems are in particular also exchangeable and coherent. A *precise exchangeable predictive system* is regularly exchangeable if and only if $p^{N-1}_{\mathscr{X}}(x_1,\dots,x_{N-1}) > 0$ for all $(x_1,\dots,x_{N-1}) \in \mathscr{X}^{N-1}$ and all finite and non-empty sets $\mathscr{X}$. This shows that regular exchangeability is a stricter requirement than exchangeability.

The term *regular* here reminds of the notion of regular extension considered by Walley in [13]. In a regularly exchangeable predictive system every predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ is the lower envelope of the predictive linear previsions $P^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$, which can be uniquely derived from the joint linear previsions $P^N_{\mathscr{X},\gamma}$ by applying Bayes's rule:

$$P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x}) = \frac{P^N_{\mathscr{X},\gamma}(fI_{\{\boldsymbol{x}\}\times\mathscr{X}^{N-n}})}{P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}^{N-n})}$$

for every gamble $f \in \mathscr{L}(\mathscr{X})$ and sample $\boldsymbol{x} \in \mathscr{X}^n$, or equivalently,

$$p^{n+1}_{\mathscr{X},\gamma}(z|\boldsymbol{x}) = \frac{p^{n+1}_{\mathscr{X},\gamma}(\boldsymbol{x},z)}{p^n_{\mathscr{X},\gamma}(\boldsymbol{x})}$$

for all $z \in \mathscr{X}$ and $\boldsymbol{x} \in \mathscr{X}^n$, because the probability $p^n_{\mathscr{X},\gamma}(\boldsymbol{x}) := P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}^{N-n})$ of the conditioning event is non-zero.

In regularly exchangeable predictive systems, the number of times

$$T_z(\boldsymbol{x}) := |\{k \in \{1,\dots,n\} : x_k = z\}|$$

that a given category z in $\mathscr{X}$ has been observed in some sample $\boldsymbol{x} \in \mathscr{X}^n$ of length $0 \le n \le N$, is of special importance. This leads us to consider the *counting map $\boldsymbol{T}_{\mathscr{X}}$* that maps samples $\boldsymbol{x}$ of length n to the $\mathscr{X}$-tuple $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ whose components are $T_z(\boldsymbol{x})$, $z \in \mathscr{X}$. $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ tells us how many times each of the elements of $\mathscr{X}$ appears in the sample $\boldsymbol{x}$, and as $\boldsymbol{x}$ varies over $\mathscr{X}^n$, $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ assumes all values in the set of *count vectors* $\mathscr{N}^n_{\mathscr{X}} := \left\{\boldsymbol{m} \in \mathbb{N}^{\mathscr{X}}_0 : \sum_{z\in\mathscr{X}} m_z = n\right\}$. It is easy to see that any two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n have the same count vector $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y})$ if and only if there is some permutation π of $\{1,\dots,n\}$ such that $\boldsymbol{y} = \pi\boldsymbol{x}$.

Proposition 1. *In any precise exchangeable predictive system σ^N, consider any finite and non-empty set $\mathscr{X}$, $0 \le n \le N-1$, and samples $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathscr{X}^n$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y})$. Then $p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(\boldsymbol{y})$ and moreover, if $p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(\boldsymbol{y}) > 0$, then also $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) = P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{y})$.*

In any regularly exchangeable predictive system, the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ only depend on the sample $\boldsymbol{x}$ through its count vector $\boldsymbol{m} = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$: for any other sample $\boldsymbol{y}$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y}) = \boldsymbol{m}$, it holds that $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{y})$ and we use the notation $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ for $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in order to reflect this. In fact, from now on we only consider predictive systems—be they regularly exchangeable or not—for which the predictive lower previsions only depend on the observed samples through their count vectors, i.e., for which the count vectors are *sufficient statistics*.

One important reason for introducing regular exchangeability, is that it allows us to prove the following inequality, which has far-reaching consequences and which shall be used in Section 5.2. We denote by $\boldsymbol{e}_z$ the count vector in $\mathscr{N}^1_{\mathscr{X}}$ whose z-component is one and all of whose other components are zero; it corresponds to the case where we have a single observation which is of a category z in $\mathscr{X}$.

Proposition 2. *In any regularly exchangeable predictive system, it holds that*

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) \ge \underline{P}^{n+1}_{\mathscr{X}}(\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)|\boldsymbol{m})$$

for all finite and non-empty sets $\mathscr{X}$, all $0 \le n \le N-2$, all $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$.

Here $\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)$ denotes the gamble on $\mathscr{X}$ that assumes the value $\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_z)$ in $z \in \mathscr{X}$. It can be checked that the above inequality is an equality for precise regularly exchangeable predictive systems. The result follows then by taking lower envelopes.

3 Representation invariance and representation insensitivity

We are ready to consider Walley's notion of representation invariance; see his IDM paper [14] for more detailed discussion and motivation. While its definition seems to be fairly involved in case of general predictive inference, we shall see that it takes on a remarkably simple and intuitive form in the more special case of immediate prediction.

Representation invariance could also, and perhaps preferably so, be called *pooling invariance*. Consider a set of categories $\mathscr{X}$, and a partition $\mathscr{S}$ of $\mathscr{X}$. Each element S of such a partition corresponds to a single new category, that consists of all the elements $x \in S$ being pooled, i.e., considered as one. Denote by $S(x)$ the unique element of the partition $\mathscr{S}$ that a category $x \in \mathscr{X}$ belongs to. Now consider a gamble f on $\mathscr{X}$ that doesn't differentiate between pooled categories, or in other words, that is constant on the elements of $\mathscr{S}$. This f can be seen as a gamble $\tilde{f}$ on the set of categories $\mathscr{S}$, such that $\tilde{f}(S(x)) := f(x)$ for all $x \in \mathscr{X}$. Similarly, with a sample $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathscr{X}^n$, there corresponds a sample $S(\boldsymbol{x}) := (S(x_1), \dots, S(x_n)) \in \mathscr{S}^n$ of pooled categories. We consider $\mathscr{S}$ as a new set of categories, and representation invariance now requires that

$$\underline{P}^{n+1}_{\mathscr{S}}(\tilde{f}|S(\boldsymbol{x})) = \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}),$$

i.e., for gambles that do not differentiate between pooled categories, it should not matter whether we consider predictive inferences for the set of original categories $\mathscr{X}$, or for the set of pooled categories $\mathscr{S}$.

We are especially interested in predictive inference where a subject starts from a state of prior ignorance. In such a state, he has no reason to distinguish between the different elements of any set of categories $\mathscr{X}$ he has chosen. How can this be expressed in terms of predictive lower previsions? Consider a permutation ϖ of the elements of $\mathscr{X}$.[2] With any gamble f on $\mathscr{X}$, there corresponds a permuted gamble $\varpi f = f \circ \varpi$. Similarly, with an observed sample $\boldsymbol{x}$ in $\mathscr{X}^n$, there corresponds a permuted sample $\varpi\boldsymbol{x} = (\varpi(x_1), \dots, \varpi(x_n))$. If a subject has no reason to distinguish between categories z and their images ϖz, this means that

$$\underline{P}^{n+1}_{\mathscr{X}}(\varpi f|\boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{X}}(f|\varpi\boldsymbol{x}).$$

We call this property *category permutation invariance*.[3]

We call *representation insensitivity* the combination of both representation invariance and category permutation invariance. 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, i.e., category set. To make this more explicit, consider two non-empty and finite sets of categories $\mathscr{X}$ and $\mathscr{Y}$, and a so-called *relabeling map* $\rho\colon \mathscr{X} \to \mathscr{Y}$ that is *onto*, i.e., such that $\mathscr{Y} = \rho(\mathscr{X}) := \{\rho(x)\colon x \in \mathscr{X}\}$. Then with any gamble f on $\mathscr{Y}$ there corresponds a gamble $\rho f := f \circ \rho$ on $\mathscr{X}$. Similarly, with an observed sample $\boldsymbol{x}$ in $\mathscr{X}^n$, there corresponds a transformed sample $\rho\boldsymbol{x} = (\rho(x_1), \dots, \rho(x_n))$ in $\mathscr{Y}^n$. *Representation insensitivity for immediate prediction then means that* $\underline{P}^{n+1}_{\mathscr{X}}(\rho f|\boldsymbol{x})$ *should be equal to* $\underline{P}^{n+1}_{\mathscr{Y}}(f|\rho\boldsymbol{x})$.

3.1 Definition and basic properties

For any gamble f on a finite and non-empty set of categories $\mathscr{X}$, its range $f(\mathscr{X}) := \{f(x)\colon x \in \mathscr{X}\}$ can again be considered as a finite and non-empty set of categories, and f itself can be considered as a relabeling map. With any $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ there corresponds a count vector $\boldsymbol{m}^f$ in $\mathscr{N}^n_{f(\mathscr{X})}$ defined by

$$m^f_r := \sum_{f(x)=r} m_x$$

for all $r \in f(\mathscr{X})$. Clearly, if $\boldsymbol{x}$ is a sample with count vector $\boldsymbol{m}$, then the relabeled sample $f\boldsymbol{x} = (f(x_1), \dots, f(x_n))$ has count vector $\boldsymbol{m}^f$. Representation insensitivity is then equivalent to the following requirement, which we take as its definition, because of its simplicity and elegance.

Definition 6 (Representation insensitivity)**.** *A predictive system* σ^N *is* representation insensitive *if for all* $0 \le n \le N-1$, *for any finite and non-empty sets* $\mathscr{X}$ *and* $\mathscr{Y}$, *for any* $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ *and* $\boldsymbol{m}' \in \mathscr{N}^n_{\mathscr{Y}}$, *and for any gambles* f *on* $\mathscr{X}$ *and* g *on* $\mathscr{Y}$ *such that* $f(\mathscr{X}) = g(\mathscr{Y})$, *the following implication holds:*

$$\boldsymbol{m}^f = \boldsymbol{m}'^g \Rightarrow \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{\mathscr{Y}}(g|\boldsymbol{m}').$$

Clearly, a predictive system σ^N is representation insensitive if and only if for all finite and non-empty sets $\mathscr{X}$, all $0 \le n \le N-1$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all $f \in \mathscr{L}(\mathscr{X})$:

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{f(\mathscr{X})}(\mathrm{id}_{f(\mathscr{X})}|\boldsymbol{m}^f), \tag{2}$$

where $\mathrm{id}_{f(\mathscr{X})}$ denotes the identity map (gamble) on $f(\mathscr{X})$. The predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ then depends on $f(\mathscr{X})$ and $\boldsymbol{m}^f$ only, and not directly on $\mathscr{X}$, f and $\boldsymbol{m}$. More explicitly, $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ *only depends on the values that* f *may assume, and on the number of times each value has been observed.*

We denote by $\Sigma^N_{\mathrm{e,ri}}$ the set of all exchangeable predictive systems that are representation insensitive. It is a subset of the class Σ^N_{e} of all exchangeable predictive systems, and

[2] This permutation ϖ of the elements of $\mathscr{X}$, or in other words of the *categories*, should be contrasted with the permutation π of the order of the observations, i.e., of the time set $\{1, \dots, N\}$, considered in Section 2.4 in order to define exchangeability.

[3] It is related to the notion of (weak) permutation invariance that two of us studied in much detail in a paper [4] dealing with general issues of symmetry in uncertainty modelling.

it inherits the order $\preceq$. Clearly, taking (non-empty) infima preserves representation insensitivity, so $\langle \Sigma^N_{\mathrm{e,ri}}, \preceq \rangle$ is a complete semi-lattice as well. We shall see in Theorem 5 that these two structures have the same bottom (the vacuous representation insensitive and exchangeable predictive system).

The remainder of this paper is devoted to the predictive systems in $\langle \Sigma^N_{\mathrm{e,ri}}, \preceq \rangle$. So we are interested in finding, and studying the properties of, predictive systems that are both exchangeable (and therefore coherent) *and* representation insensitive. We believe performing such a study to be quite important, and we here report on our first attempts.

3.2 The lower probability function

With any predictive system σ^N, we can associate a map φ_{σ^N} that is defined on the subset $\{(n,m)\colon 0 \le m \le n \le N-1\}$ of $\mathbb{N}^2_0$ by

$$\varphi_{\sigma^N}(n,m) := \underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}}|n-m,m).$$

Why this map is important, becomes clear if we look at predictive systems that are representation insensitive. Consider any proper event $\emptyset \neq A \subsetneq \mathscr{X}$, then it follows by applying Equation (2) with $f = I_A$, that

$$\begin{aligned}\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) &= \underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}}|n-m_A,m_A)\\ &= \varphi_{\sigma^N}(n,m_A) \qquad (3)\end{aligned}$$

where $m_A := \sum_{z\in A} m_z$. So we see that in a representation insensitive predictive system, the lower probability of observing an event (that is neither considered to be impossible nor necessary) does not depend on the embedding set $\mathscr{X}$ nor on the event itself, but only on the total number of previous observations n, and on the number of times m that the event has been observed before, and is given by $\varphi_{\sigma^N}(n,m)$. Something similar holds of course for the upper probability of observing a non-trivial event. Indeed, by conjugacy,

$$\begin{aligned}\overline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) &= 1-\underline{P}^{n+1}_{\mathscr{X}}(A^c|\boldsymbol{m}) = 1-\varphi_{\sigma^N}(n,m_{A^c})\\ &= 1-\varphi_{\sigma^N}(n,n-m_A). \qquad (4)\end{aligned}$$

This property of representation insensitive predictive systems is reminiscent of *Johnson's sufficientness postulate* [9] (we use Zabell's terminology [17]), which requires that the probability that the next observation will be a category x is a function $f_x(n,m_x)$ that depends only on the category x itself, on the number of times m_x that this category has been observed before, and on the total number of previous observations n. Representation insensitivity is stronger, because it entails that the function φ_{σ^N} that 'corresponds to' the f_x is the same for all categories x in all possible finite sets and non-empty $\mathscr{X}$.

We call φ_{σ^N} the *lower probability function* of the predictive system σ^N. We shall simply write φ instead of φ_{σ^N}, whenever it is clear from the context which predictive system we are talking about. Let us give a number of interesting properties for the lower probability function φ associated to a representation insensitive and coherent predictive system σ^N.

Proposition 3. *Let $N > 0$ and let σ^N be a representation insensitive and coherent predictive system with lower probability function φ. Then*

1. *φ is $[0,1]$-bounded:* $0 \le \varphi(n,k) \le 1$ *for all* $0 \le k \le n \le N-1$.
2. *φ is super-additive in its second argument:* $\varphi(n,k+\ell) \ge \varphi(n,k)+\varphi(n,\ell)$ *for all non-negative integers n, k and ℓ such that* $k+\ell \le n \le N-1$.
3. $\varphi(n,0) = 0$ *for all* $0 \le n \le N-1$.
4. $\varphi(n,k) \ge k\varphi(n,1)$ *for* $1 \le k \le n \le N-1$, *and* $0 \le n\varphi(n,1) \le 1$ *for* $1 \le n \le N-1$.
5. *φ is non-decreasing in its second argument:* $\varphi(n,k+1) \ge \varphi(n,k)$ *for* $0 \le k < n \le N-1$.

If σ^N is moreover regularly exchangeable, then

6. $\varphi(n+1,k)+\varphi(n,k)[\varphi(n+1,k+1)-\varphi(n+1,k)] \le \varphi(n,k)$ *for* $0 \le k \le n \le N-2$.
7. *φ is non-increasing in its first argument:* $\varphi(n+1,k) \le \varphi(n,k)$ *for* $0 \le k \le n \le N-2$.
8. $\varphi(n,1) \ge \varphi(n+1,1)[1+\varphi(n,1)]$ *for* $1 \le n \le N-2$.
9. *Suppose that $\varphi(n,1) > 0$ and define $s_n := \frac{1}{\varphi(n,1)} - n$ for $1 \le n \le N-1$.*[4] *Then $s_n \ge 0$, s_n is non-decreasing and $\varphi(n,1) = 1/(s_n+n)$.*

In particular, these results, together with Equations (3) and (4), allow us to draw interesting and intuitively appealing conclusions about predictive lower and upper probabilities, which are valid in any representation insensitive and coherent predictive system: (i) the lower probability of observing an event that hasn't been observed before is zero, and the upper probability of observing an event that has always been observed before is one [Proposition 3.3]; and (ii) if the number of observations remains fixed, then both the lower and the upper probability of observing an event again do not decrease if the number of times the event has already been observed increases [Proposition 3.5]. In predictive systems that are moreover regularly exchangeable, we also see that (iii) if the number of times an event has been observed remains the same as the number of observations increases, then the lower probability for observing the event again does not increase [Proposition 3.7].

[4] This s_n will later, in Section 5.2 turn out to be a constant (independent of the number of observations n) under special additional assumptions, and will play the rôle of the hyper-parameter s in the ID(M)M.

When the predictive system consists solely of families of predictive linear previsions (apart from predictive lower previsions for dealing with zero previous observations, see Section 4), we can use the additivity of linear previsions, instead of the mere super-additivity of coherent lower previsions used previously, to get stronger versions of parts of Proposition 3[5]. Such predictive systems will be characterised in Theorem 6 further on.

Corollary 4. *Consider a representation insensitive and coherent predictive system σ^N, with a lower probability function φ, and such that all the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ for $0 < n \leq N-1$ are linear previsions. Then for all $0 < n \leq N-1$ and all $k, \ell \geq 0$ such that $k+\ell \leq n$:*

1. $\varphi(n, k+\ell) = \varphi(n,k) + \varphi(n,\ell)$.
2. $\varphi(n,k) = k\varphi(n,1)$.

4 Are there representation insensitive exchangeable predictive systems?

We don't know yet if there are any predictive systems that are both representation insensitive and exchangeable. We remedy this situation here by establishing the existence of two 'extreme' types of representation insensitive and exchangeable predictive systems, one of which is also regularly exchangeable.

Consider, for any predictive system σ^N that is both representation insensitive and exchangeable, the predictive lower previsions for $n = 0$. These are actually unconditional lower previsions $\underline{P}^1_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X})$, modelling our beliefs about the first observation X_1, i.e., when no observations have yet been made. It follows right away from Proposition 3 and Equations (3) and (4) that for any proper subset A of $\mathscr{X}$, $\underline{P}^1_{\mathscr{X}}(A) = \varphi(0,0) = 0$. Since $\underline{P}^1_{\mathscr{X}}$ is assumed to be a (separately) coherent lower prevision, it follows that $\underline{P}^1_{\mathscr{X}}(f) = \min f$, for any gamble f on $\mathscr{X}$. So *all the $\underline{P}^1_{\mathscr{X}}$ in a representation insensitive and exchangeable predictive system must be so-called vacuous lower previsions.*[6] This means that there is no choice for the first predictions. It also means that it is impossible to achieve representation insensitivity in any precise predictive system (but see Theorem 6 for a predictive system that comes close).

This leads us to consider the so-called *vacuous* predictive system ν^N where all predictive previsions are vacuous: for all $0 \leq n \leq N-1$, all finite and non-empty sets of categories $\mathscr{X}$, all $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$, $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \min f$.

Theorem 5. *The vacuous predictive system ν^N is regularly exchangeable and representation insensitive. It is the bottom (smallest element) of the complete semilattice $\langle \Sigma^N_{\mathrm{e,ri}}, \preceq \rangle$. Its lower probability function is given by $\varphi(n,m) = 0$ for $0 \leq m \leq n \leq N-1$.*

In the vacuous predictive system the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ are all vacuous, and therefore do not depend on the number of observations n, nor on the observed count vectors $\boldsymbol{m}$. A subject who is using the vacuous predictive system is not learning anything from the observations. Representation insensitivity and (regular) exchangeability do not guarantee that we become more committal as we have more information at our disposal. Indeed, with the vacuous predictive system, whatever our subject has observed before, he always remains fully uncommittal. If we want a predictive system where something is really being learned from the data, it seems we need to make some 'leap of faith', and add something to our assessments that is not a mere consequence of exchangeability and representation insensitivity.

So are there less trivial examples of exchangeable and representation insensitive predictive systems? We must make the vacuous choice for $n = 0$, but is there, for instance, a way to make the predictive lower previsions *precise*, or linear, for $n > 0$? The following theorem tells us there is only one such exchangeable and representation insensitive predictive system.

Theorem 6. *Consider a predictive system where for any $0 < n \leq N-1$ all the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ are actually linear previsions $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$. If this predictive system is representation insensitive, then*

$$P^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \sum_{z \in \mathscr{X}} f(z) \frac{m_z}{n} \tag{5}$$

for all $0 < n \leq N-1$, all finite and non-empty sets of categories $\mathscr{X}$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$. For its lower probability function φ, we then have $\varphi(n,k) = \frac{k}{n}$ for all $0 \leq k \leq n$ and $n > 0$. Moreover, the predictive previsions given by Equation (5)*, together with the vacuous lower previsions for $n = 0$, constitute a representation insensitive and exchangeable (but not regularly so) predictive system π^N.*

We call the predictive system π^N described in Theorem 6 the *Haldane* predictive system. The name refers to the fact that a Bayesian inference model with a multinomial likelihood function using Haldane's (improper) prior (see, e.g., Jeffreys [8, p. 123]) leads to these predictive previsions for $n > 0$.

It is a consequence of Walley's Marginal Extension Theorem [13, Section 6.7.3] that for any finite and non-empty $\mathscr{X}$, the only joint lower prevision on $\mathscr{L}(\mathscr{X}^N)$

[5] Note that the equalities in this corollary will also hold for some non-linear predictive systems, such as the mixing ones we shall consider in Section 5

[6] This result was proven, in another way, by Walley [13, Section 5.5.1], when he argued that his Embedding and Symmetry Principles under coherence only leave room for the vacuous lower prevision. When there are no prior observations ($n = 0$), the Embedding Principle is related to representation invariance, and the Symmetry Principle with what we have called category permutation invariance.

that is coherent with the Haldane predictive $\mathscr{X}$-family is given by $\underline{P}^N_{\mathscr{X}}(f) = \min_{z\in\mathscr{X}} f(z,\ldots,z)$. This implies that the Haldane predictive system is not regularly exchangeable: any dominating precise exchangeable predictive system satisfies $p^{N-1}_{\mathscr{X}}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathscr{X}^{N-1}$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{m} \neq (N-1)\boldsymbol{e}_z$ for all $z \in \mathscr{X}$, and for any such $\boldsymbol{x}$, the requirements for regular exchangeability cannot be satisfied.

The Haldane predictive system only seems to be coherent with a joint lower prevision $\underline{P}^N_{\mathscr{X}}$ which expresses that our subject is certain that all variables X_k will assume *the same value*, but where he is completely ignorant about what that common value is. This is related to another observation: we deduce from Proposition 3.3 that in the Haldane predictive system, when $n > 0$ then not only the lower probability but also the upper probability of observing an event that hasn't been observed before is zero! This models that a subject is practically certain (because prepared to bet at all odds on the fact) that any event that hasn't been observed in the past will not be observed in the future either. The *sampling prevision* $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ for a gamble f in this predictive system is the expectation of f with respect to the observed (sampling) probability distribution on the set of categories. The Haldane predictive system is too strongly tied to the observations, and does not allow us to make 'reasonable' inferences in a general context.

5 Mixing predictive systems

So we have found two extreme representation insensitive and exchangeable predictive systems, both of which are not very useful: the first, because it doesn't allow us to learn from past observations, and the second, because its inferences are too strong and we seem to infer too much from the data. A natural question then is: can we find 'intermediate' representation insensitive and exchangeable predictive systems whose behaviour is stronger than the vacuous predictive system and weaker than the Haldane predictive system? The first idea that comes to mind, is to look at convex mixtures. Let us, therefore, consider a finite sequence ε, of N numbers $\varepsilon_n \in [0,1]$, $0 \leq n \leq N-1$, and study the *mixing predictive system* σ^N_ε whose predictive lower previsions are given by

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \varepsilon_n S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) + (1-\varepsilon_n)\min f, \qquad (6)$$

for all $0 \leq n \leq N-1$, all finite and non-empty sets of categories $\mathscr{X}$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$. As $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ is only defined for $n > 0$, and since representation insensitivity and coherence require that $\underline{P}^1_{\mathscr{X}}$ should be vacuous, we always let $\varepsilon_0 = 0$ implicitly. We call any such sequence ε a *mixing sequence*, and we denote by φ_ε the lower probability function of the corresponding mixing predictive system σ^N_ε.

We are mainly interested in finding mixing predictive systems that are representation insensitive and (regularly) exchangeable. The following proposition tells us that the only real issue lies with exchangeability.

Proposition 7. *For any mixing sequence ε, the predictive system σ^N_ε is still representation insensitive. Moreover, let $0 \leq k \leq n \leq N-1$. Then $\varphi_\varepsilon(n,k) = \varepsilon_n\frac{k}{n}$, and if $\varepsilon_n > 0$ then $s_n = n\frac{1-\varepsilon_n}{\varepsilon_n}$ and $\varepsilon_n = \frac{n}{n+s_n}$. In particular $\varphi_\varepsilon(n,1) = \varepsilon_n/n$ is the lower probability of observing a non-trivial event that has been observed once before in n trials, $\varepsilon_n = n\varphi_\varepsilon(n,1)$ is the lower probability $\varphi_\varepsilon(n,n)$ of observing a non-trivial event that has always been observed before (n out of n times), and $s_n = \frac{1-\varphi_\varepsilon(n,n)}{\varphi_\varepsilon(n,1)}$ is the ratio of the upper probability of observing an event that has never been observed before to the lower probability of observing an event that has been observed once before, in n trials.*

We have already argued that in order to get away from making vacuous inferences, and in order to be able to learn from observations, we need to make some 'leap of faith' and go beyond merely requiring exchangeability and representation insensitivity. *One of the simplest ways to do so, is to specify the numbers $\varphi(n,1)$ for $n = 1,\ldots,N-1$, or in other words, to specify, beforehand, the lower probability of observing any non-trivial event that has been observed only once in n trials.* We can then ask for the most conservative representation insensitive predictive system that exhibits these lower probabilities. The following theorem tells us that mixing predictive systems play this part.

Theorem 8. *Consider $N > 0$ and a mixing sequence ε. Let σ^N be a representation insensitive coherent predictive system such that its associated lower probability function φ satisfies*

$$\varphi(n,1) \geq \varphi_\varepsilon(n,1) = \varepsilon_n/n$$

for all $0 < n \leq N-1$. Then $\sigma^N_\varepsilon \preceq \sigma^N$.

Mixing predictive systems have a special part in this theory, because they are quite simple, and in some sense most conservative. They are quite simple because all that is needed to specify them is the values $\varphi(n,1)$ of the lower probability function, or in other words, the lower probabilities that an event will occur that has been observed once in n observations. They are the most conservative coherent and representation insensitive predictive systems with the given values for $\varphi(n,1)$. In the following subsections we shall see that there are mixing predictive systems with a non-trivial mixing sequence ε that are also regularly exchangeable, and we derive a necessary condition on the mixing sequence ε for this to be the case.

5.1 The regular exchangeability of mixing predictive systems

Consider any mixing sequence ε and the corresponding mixing predictive system σ^N_ε. For the corresponding lower probability function φ_ε it holds by Proposition 7 that

$\varphi_\varepsilon(n,k) = \varepsilon_n \frac{k}{n}$; if we substitute this in the inequality of Proposition 3.8 we see that it is necessary for regular exchangeability that

$$\frac{\varepsilon_n}{n} \geq \frac{\varepsilon_{n+1}}{n+1}\left(1+\frac{\varepsilon_n}{n}\right), \quad n=1,\dots,N-1. \tag{7}$$

If one ε_n is zero, then all of the subsequent ε_{n+k} are zero as well: if inferences are vacuous after $n > 0$ observations, they should also remain vacuous after subsequent ones. Or, to put it more boldly, in regularly exchangeable mixing predictive systems, if we are going to learn at all from observations, we have to start doing so from the first observation.

5.2 Predictive inferences for the IDMM

It is of particular interest to investigate for which types of mixing predictive systems, or in other words, for which mixing sequences ε, we generally have an equality rather than only an inequality in the condition of Proposition 2, i.e., for which

$$\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m}) = \underline{P}_{\mathscr{X}}^{n+1}(\underline{P}_{\mathscr{X}}^{n+2}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)|\boldsymbol{m}), \tag{8}$$

for all finite and non-empty $\mathscr{X}$, all $0 \leq n \leq N-1$, all $\boldsymbol{m} \in \mathscr{N}_{\mathscr{X}}^n$ and all gambles f on $\mathscr{X}$, where the predictive lower previsions $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{m})$ are given by Equation (6). Using the definition of $S_{\mathscr{X}}^{n+1}(f|\boldsymbol{m})$, and the coherence of $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{m})$ we find, after some rearranging, that Equation (8) holds if and only if

$$\frac{\varepsilon_n}{n} = \frac{\varepsilon_{n+1}}{n+1}\left(1+\frac{\varepsilon_n}{n}\right), \quad n=1\dots,N-1,$$

i.e., we have the equality in (7). Clearly, one ε_n is zero if and only if all of them are, which leads to the vacuous predictive system v^N. We already know this vacuous system to be regularly exchangeable (and representation insensitive). If we assume on the other hand that $\varepsilon_n > 0$ for $n = 1,\dots,N$, and let $\zeta_n := n/\varepsilon_n = n + s_n \geq 1$, then the above equality can be rewritten as $\zeta_{n+1} = \zeta_n + 1$, which implies that there is some $s \geq 0$ such that $\zeta_n = n + s$, or equivalently, $s_n = s$ and consequently, $\varepsilon_n = \frac{n}{n+s}$, and

$$\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m}) = \frac{n}{n+s} S_{\mathscr{X}}^{n+1}(f|\boldsymbol{m}) + \frac{s}{n+s}\min f \tag{9}$$

for $n = 0,1,\dots,N-1$. The predictive lower previsions in Equation (9) are precisely the ones that can be associated with the so-called Imprecise Dirichlet-Multinomial Model (or IDMM) with hyper-parameter s [15, Section 4.1]. We call mixing predictive systems of this type IDMM-*predictive systems*. The vacuous predictive system corresponds to letting $s \to \infty$.

Theorem 9. *The vacuous predictive system, and the* IDMM-*predictive systems for $s > 0$ are regularly exchangeable and representation insensitive, and they are the only mixing predictive systems for which the equality* (8) *holds.*

Among the mixing predictive systems, the ones corresponding to the IDMM are also special in another way. which points to a quite peculiar, but intuitively appealing, property of predictive inferences produced by the IDMM. Indeed, assume that in addition to observing a count vector $\boldsymbol{m}$ of n observations, we know in some way that the $(n+1)$-th observation will belong to a proper subset A of $\mathscr{X}$—we might suppose for instance that the observation X_{n+1} has been made, but that it is imperfect, and only allows us to conclude that $X_{n+1} \in A$. Then we can ask what the updated beliefs are, i.e., what $\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m},A)$ is. Since $\underline{P}_{\mathscr{X}}^{n+1}(A|\boldsymbol{m}) = \varepsilon_n m_A/n > 0$ if and only if $m_A > 0$ and $\varepsilon_n > 0$, let us assume that indeed $m_A > 0$ and $\varepsilon_n > 0$, in which case the requirements of coherence allow us to determine $\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m},A)$ uniquely, using the so-called Generalised Bayes Rule [13, Section 6.4]. This implies that $\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m},A)$ is then the unique real μ such that

$$\underline{P}_{\mathscr{X}}^{n+1}(I_A(f-\mu)|\boldsymbol{m}) = 0.$$

We now have the following characterisation of IDMM-predictive systems.

Theorem 10 (Specificity)**.** *The* IDMM-*predictive systems with $s > 0$ are the only mixing predictive systems with all $\varepsilon_n > 0$, $n = 1,\dots,N-1$ that satisfy*

$$\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m},A) = \underline{P}_{A}^{m_A+1}(f_A|\boldsymbol{m}_A) \tag{10}$$

for all $n = 1,\dots,N-1$, all $\boldsymbol{m} \in \mathscr{N}_{\mathscr{X}}^n$, all gambles f on $\mathscr{X}$ and all proper subsets A of $\mathscr{X}$ such that $m_A > 0$.

We have denoted by f_A the restriction of the gamble f to the set A, by $\boldsymbol{m}_A$ the A-tuple obtained from $\boldsymbol{m}$ by dropping the components that correspond to elements outside A. The sum of the components of $\boldsymbol{m}_A$ is m_A.

This so-called *specificity* property of inferences characterised by Equation (10) is quite peculiar. Suppose that you have observed n successive outcomes, leading to a count vector $\boldsymbol{m}$. If you know in addition that $X_{n+1} \in A$, then Equation (10) tells you that *the updated value $\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{m},A)$ is the same as the one you would get by discarding all the* previous *observations producing values outside A, and in effect only retaining the m_A observations that were inside A!* Knowing that the $(n+1)$-th observation belongs to A allows you to ignore all the previous observations that happened to lie outside A. This is intuitively appealing, because it means that if you know that the outcome of the next observation belongs to A, only the related behaviour (the values of f on A and the previous observations of this set) matters for your prediction.

The name 'specificity' for this property was suggested to us by Jean-Marc Bernard. In one of his papers [1], he calls 'specific' any type of inference that has this particular property.

6 Conclusions

More work is needed in order to be able to draw a reasonably complete picture of the issue of representation insensitivity in predictive systems. Indeed, while doing research for this paper, we came across a multitude of questions that we haven't yet been able to answer, and we list only a few of them here.

(i) Are there (regularly) exchangeable and representation insensitive predictive systems that are not mixing predictive systems?

(ii) Related questions are: are there (regularly) exchangeable and representation insensitive predictive systems that, unlike the mixing systems, are not completely determined by the probabilities $\varphi(n,1)$ of observing an event that has been observed only once before in n observations; are there such predictive systems whose behaviour on gambles, unlike that of mixing systems, is not completely determined by the lower probability function φ; and are there such predictive systems whose lower probability function φ, unlike that of mixing systems, is not additive in the sense that $\varphi(n,k+\ell) = \varphi(n,k) + \varphi(n,\ell)$?

(iii) Are there (regularly) exchangeable and representation insensitive mixing predictive systems that are not of the IDMM-type? And if so,

(iv) are there (regularly) exchangeable, representation insensitive non-mixing predictive systems that satisfy Equation (10)?

(v) Can we arrive at stronger conclusions if we consider that the observations X_n make up an infinite exchangeable sequence?

(vi) Can more definite answers be given if we consider the general, rather than the immediate, prediction problem?

Acknowledgements

We thank Jean-Marc Bernard, Frank Coolen, Thomas Augustin, and the reviewers for useful discussions and comments.

References

[1] J.-M. Bernard. Bayesian analysis of tree-structured categorized data. *Revue Internationale de Systémique*, 11:11–29, 1997.

[2] R. Carnap. *The continuum of inductive methods*. The University of Chicago Press, 1952.

[3] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, Cambridge, 1990.

[4] G. de Cooman and E. Miranda. Symmetry of models versus models of symmetry. In W. L. Harper and G. R. Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.* King's College Publications, 2007.

[5] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937.

[6] B. de Finetti. *Teoria delle Probabilità*. Einaudi, Turin, 1970.

[7] B. de Finetti. *Theory of Probability*. John Wiley & Sons, Chichester, 1974–1975. English translation of [6], two volumes.

[8] H. Jeffreys. *Theory of Probability*. Oxford Classics series. Oxford University Press, 1998. Reprint of the third edition (1961), with corrections.

[9] W. E. Johnson. *Logic, Part III. The Logical Foundations of Science*. Cambridge University Press, 1924. Reprinted by Dover Publications in 1964.

[10] P.-S. Laplace. *Philosophical Essay on Probabilities*. Dover Publications, 1951. English translation of [11].

[11] P.-S. Laplace. *Essai philosophique sur les probabilités*. Christian Bourgois Éditeur, 1986. Reprinted from the fifth edition (1825).

[12] E. Miranda and G. de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 2007. Accepted for publication.

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

[14] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.

[15] P. Walley and J.-M. Bernard. Imprecise probabilistic prediction for categorical data. Technical Report CAF-9901, Laboratoire Cognition et Activitées Finalisés, Université de Paris 8, January 1999.

[16] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Reprinted in a revised form in the International Journal of Approximate Reasoning, 44(3), 366-383, 2007.

[17] S. L. Zabell. W. E. Johnson's "sufficientness" postulate. *The Annals of Statistics*, 10:1090–1099, 1982.