SYSTeMS Research Group
Ghent University
Gert.deCooman@UGent.be
Erik.Quaeghebeur@UGent.be

Immediate prediction under exchangeability & representation insensitivity

Gert de Cooman, Enrique Miranda & Erik Quaeghebeur

Department of Statistics
and Operations Research
Rey Juan Carlos University
Enrique.Miranda@urjc.es

1 The setting

Sampling: A subject makes a fixed number $N>0$ of successive observations, represented by *random variables* $X_1,\dots,X_N$. For example, when drawing coloured balls without replacement from an urn, X_k designates the unknown colour of the k-th ball.

Immediate prediction: The subject in some way uses zero or more observations $X_1,\dots,X_n$ made previously (so n belongs to $\{0,1,\dots,N-1\}$), to predict, or make inferences about, the value of the *next* observation X_{n+1}.

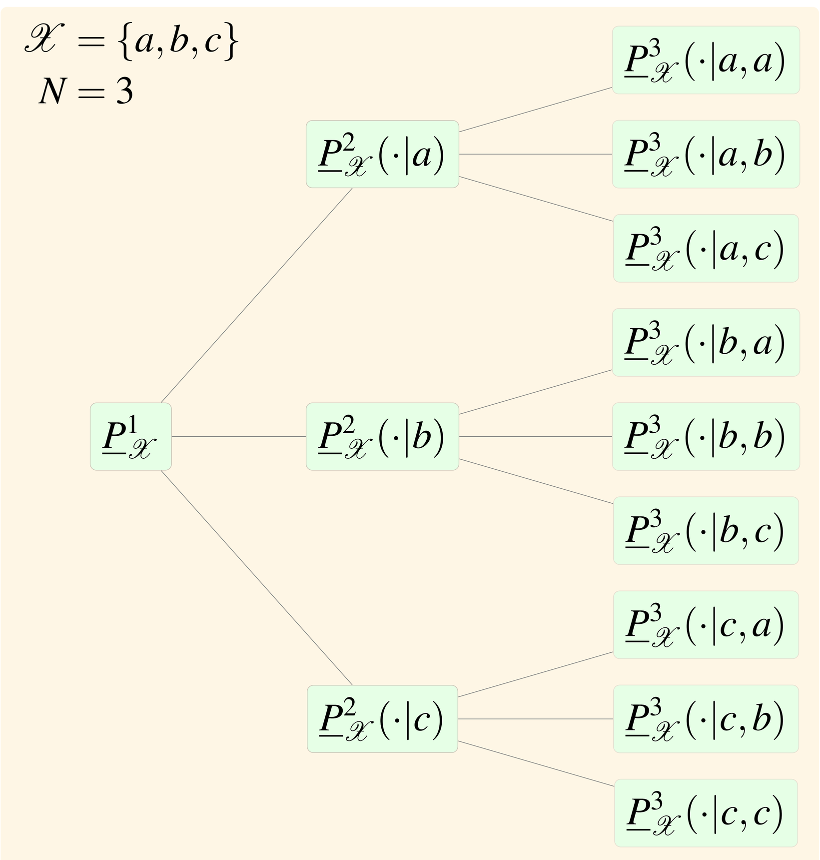
Families of predictive lower previsions: The subject can determine, beforehand, a finite and non-empty set $\mathscr{X}$ of possible values, or *categories*, for the random variables.

For each n and each sequence $\boldsymbol{x}=(x_1,\dots,x_n)$ in $\mathscr{X}^n$, she can give a *predictive lower prevision* $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for X_{n+1}, given the values $(X_1,\dots,X_n)=(x_1,\dots,x_n)=\boldsymbol{x}$ of the previous observations. It is defined on the set of all gambles f on $\mathscr{X}$.

$\mathscr{X}=\{a,b,c\}$ $N=3$	Let $f(a)=1,\ f(b)=3,\ f(c)=-2$, then, e.g., $\underline{P}^3_{\mathscr{X}}(f\vert c,a)=-\frac{1}{2}$.
$n=2$ $\boldsymbol{x}=(c,a)$	Let $A=\{a,c\}$, then, e.g., $\underline{P}^3_{\mathscr{X}}(A\vert c,a)=\frac{4}{5}$.

An $\mathscr{X}$-*family* $\sigma^N_{\mathscr{X}}$ of predictive lower previsions is the set formed for all possible observations:

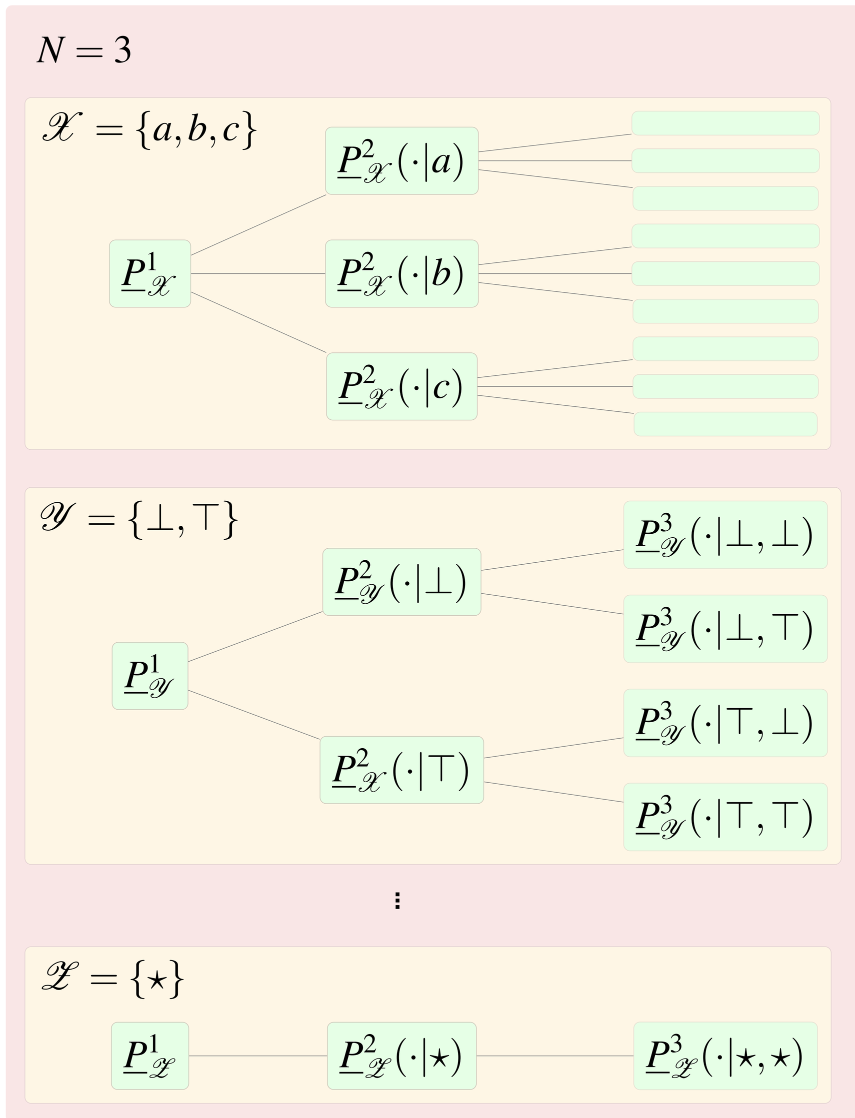
$$\sigma^N_{\mathscr{X}}:=\left\{\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})\colon \boldsymbol{x}\in\mathscr{X}^n \text{ and } n=0,1,\dots,N-1\right\}.$$



Precise predictive families are those that only contain predictive *linear* previsions $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$. With each of these, there corresponds a predictive probability mass function. They in turn allow us, using Bayes's rule, to find the unique joint probability mass functions $p^n_{\mathscr{X}}$ on $\mathscr{X}^n$ and the corresponding *joint linear prevision* $P^N_{\mathscr{X}}$, which models beliefs about the values that the random variables $(X_1,\dots,X_N)$ assume *jointly* in $\mathscr{X}^N$.

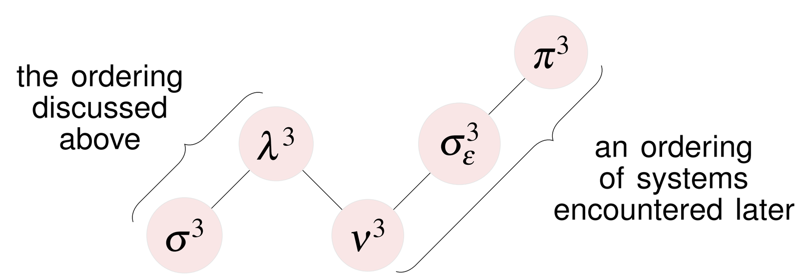
Systems of predictive lower previsions: The inferences or predictions of a predictive $\mathscr{X}$-family might depend on the actual choice of $\mathscr{X}$ made. So we let our subject consider predictive families for all conceivable choices of $\mathscr{X}$. We collect these predictive families in a *system* σ^N of predictive lower previsions:

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



Precise predictive systems are those that only contain precise predictive families.

Predictive systems can be *partially ordered*: The system σ^N is *more conservative* than the system λ^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.



A collection $\left\{\sigma^N_\gamma\colon \gamma\in\Gamma\right\}$ of predictive systems may have an infimum with respect to this partial order. Whenever it exists, this infimum system σ^N can be seen as a *lower envelope*: each of its predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ is defined as the lower envelope $\inf_{\gamma\in\Gamma}\underline{P}^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$ of the predictive lower previsions in the predictive systems σ^N_γ.

2 Requirements & Assumptions

Coherence: Coherence is a requirement on the individual predictive lower previsions.

A predictive system is called *coherent* if it is the lower envelope of a collection of precise predictive systems. This is equivalent to requiring that all the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in the system should be separately coherent.

(Regular) exchangeability: Exchangeability is an assumption about a family of predictive lower previsions.

A precise predictive system is *exchangeable* if all the associated joint linear previsions $P^N_{\mathscr{X}}$ are exchangeable, i.e., invariant under permutation of the random variables $X_1,\dots,X_N$.

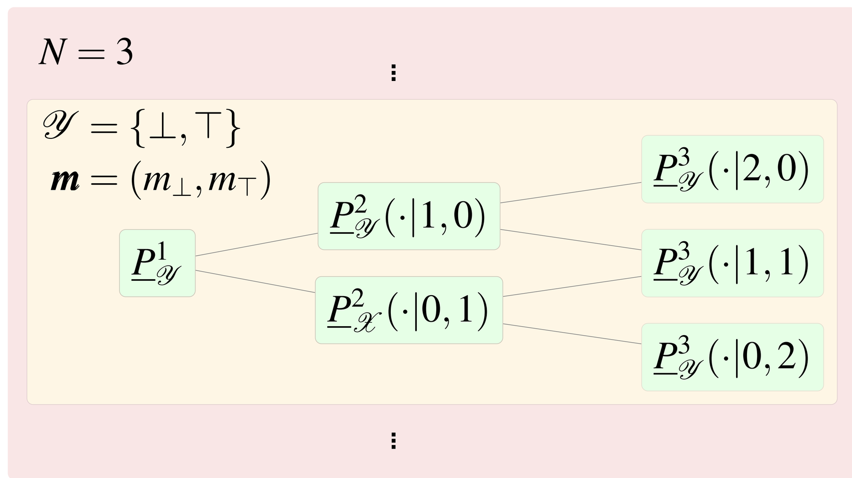
A general predictive system is called *exchangeable* if it is the lower envelope of a collection $\left\{\sigma^N_\gamma\colon \gamma\in\Gamma\right\}$ of exchangeable precise predictive systems. It is *regularly exchangeable* if all predictive linear previsions $P^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$ in each of these systems σ^N_γ can be uniquely derived from the joint linear prevision $P^N_{\mathscr{X},\gamma}$ by applying Bayes's rule. (For this, the joint mass functions $p^n_{\mathscr{X},\gamma}$ should be strictly positive for $n<N$.)

3 Some results

From sequences of observations to count vectors: In any regularly exchangeable predictive system, the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ only depend on the sequence of observations $\boldsymbol{x}$ through its *count vector* $\boldsymbol{m}\in\mathscr{N}^n_{\mathscr{X}}$, with

$$m_z:=|\{k\in\{1,\dots,n\}\colon x_k=z\}|,$$
$$\mathscr{N}^n_{\mathscr{X}}:=\left\{\boldsymbol{m}\in\mathbb{N}_0^{\mathscr{X}}\colon \textstyle\sum_{z\in\mathscr{X}}m_z=n\right\}.$$

All predictive lower previsions for given sequences with the same count vector $\boldsymbol{m}$ can therefore be written as $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$.



So, for regularly exchangeable predictive systems, count vectors are a *sufficient statistic*. From now on, we only consider (possibly non-exchangeable) predictive systems for which this is the case.

A useful (in)equality In any regularly exchangeable predictive system, it holds for all gambles f that

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

where $n\leq N-2$ and $\boldsymbol{e}_x\in\mathscr{N}^1_{\mathscr{X}}$ for $x\in\mathscr{X}$ such that, using the Kronecker delta, $(\boldsymbol{e}_x)_z=\delta_{xz}$. For precise regularly exchangeable predictive systems, this becomes a 'useful equality'.

4 Some more requirements

Representation insensitivity: Representation insensitivity is a requirement that works between predictive lower previsions for the same number of observations.

It comprises three invariance requirements:

- *pooling invariance*: inferences that do not depend on the distinction between some categories should stay the same when those categories are pooled;
- *renaming invariance*: apart from avoiding confusion, the names of the categories should not matter;
- *category permutation invariance*: in a state of prior ignorance, which we consider here, the subject has no reason to distinguish between the categories, so the inferences should be invariant under a permutation of them.

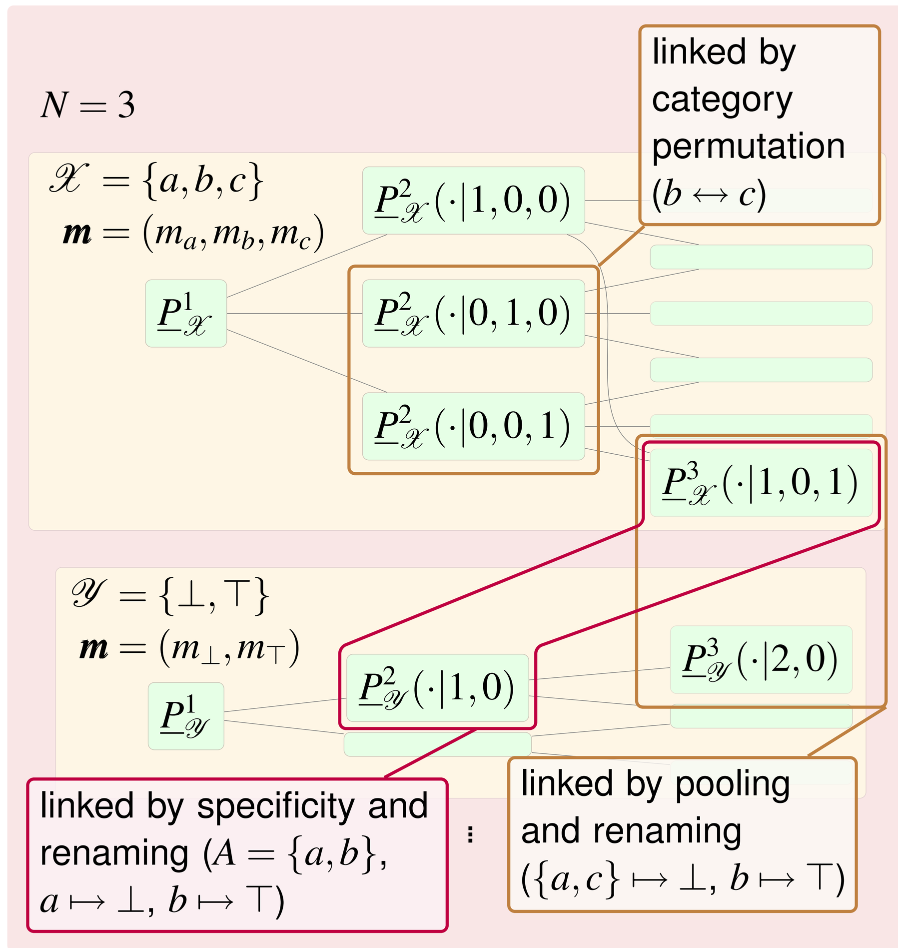
Combining these, we can say a predictive system is *representation insensitive* if for all n, for any category sets $\mathscr{X}$ and $\mathscr{Y}$, for any $\boldsymbol{m}\in\mathscr{N}^n_{\mathscr{X}}$ and $\boldsymbol{m}'\in\mathscr{N}^n_{\mathscr{Y}}$, and for any gambles f on $\mathscr{X}$ and g on $\mathscr{Y}$ with identical ranges, the following holds:

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

with $m^f_r:=\sum_{f(x)=r}m_x$. This means $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ only depends on the values that f may assume, and on the number of times each value has been observed:

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

where $\mathrm{id}_{f(\mathscr{X})}$ is the identity map on the range of f.



Specificity (optional): Specificity is a requirement that works between predictive lower previsions for a different number of observations related by pooling.

An exchangeable predictive system is *specific* if for all gambles f and all non-trivial events A in $\mathscr{X}$ containing a non-zero number m_A of observations, it holds that

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

where f_A and $\boldsymbol{m}_A$ are the restriction of f and $\boldsymbol{m}$ to A. So, knowing that the $(n+1)$-th observation belongs to A allows you to ignore all the previous observations that lie outside A.

5 More results

The lower probability function: With any predictive system we associate a map φ defined for all n and $k\leq n$ by

$$\varphi(n,k):=\underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}}|n-k,k).$$

For representation insensitive systems it fully characterizes all predictive lower probabilities (cfr. *Johnson's sufficientness postulate*) and is therefore called the *lower probability function*; to wit, let A be some event, and m_A the associated number of observations, then

$$\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m})=\underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}}|n-m_A,m_A)=\varphi(n,m_A).$$

It allows us to draw intuitively appealing conclusions, which are valid in any coherent representation insensitive system:

(i) the lower/upper probability of observing an event that has not/always been observed before is zero/one;

(ii) if n remains fixed, then both the lower and upper probability of observing A again do not decrease if m_A increases;

(iii) in systems that are also regularly exchangeable: if m_A remains the same as n increases, then the lower probability for observing A again does not increase.

Some representation insensitive exchangeable systems: To start: all the $\underline{P}^1_{\mathscr{X}}$ in a representation insensitive and exchangeable predictive system must be vacuous.

A subject that is too conservative to learn uses the regularly exchangeable *vacuous predictive system* ν^N. All its predictive lower previsions are vacuous, so $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}):=\min f$.

A subject that believes that categories unobserved in the past remain so in the future, uses the (not regularly) exchangeable *Haldane predictive system* π^N. For $n>0$, all its predictive previsions are linear and strongly tied to the observations:

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

Other systems can be formed as convex mixtures of the two extreme ones above. We define *mixing predictive systems* σ^N_ε with a $[0,1]$-bounded *mixing sequence* ε of length N and

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

note that implicitly $\varepsilon_0=0$. Representation insensitivity is retained after mixing; a sufficient condition for regular exchangeability is the reformulated 'useful inequality' $\frac{\varepsilon_n}{n}\geq\frac{\varepsilon_{n+1}}{n+1}\left(1+\frac{\varepsilon_n}{n}\right)$.

The lower probability function of a mixing system is given by $\varphi(n,k)=\varepsilon_n\frac{k}{n}$. So, as $\varepsilon_n=n\varphi(n,1)$, a mixing system can be defined by specifying the lower probability of observing any non-trivial event that has been observed once in n trials; it is the most conservative system with these lower probabilities.

The imprecise Dirichlet-Multinomial model: Any mixing system that is specific or for which the 'useful equality' holds, is uniquely characterized by some $s>0$ such that $\varepsilon_n=\frac{n}{n+s}$ and thus

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

This regularly exchangeable representation insensitive predictive system is related to the imprecise Dirichlet-Multinomial model with hyper-parameter s.

Selected references

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.*, pages 67–149. 2007.

P. Walley and J.-M. Bernard. Imprecise probabilistic prediction for categorical data. Technical Report CAF-9901, Université de Paris 8, 1999.

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

Acknowledgements

Jean-Marc Bernard, Frank-Coolen, Thomas Augustin, and the reviewers.