Bayesian Plinko

Study notes

Alex, Luca, Yasser, Britt, James

14 April 2018; updated 3 July 2018

We are human, after all
Much in common, after all
(Daft Punk 2005a)

1 Remarks, comments, thoughts on the project

1.1 [Luca] What goes on in a participant's head? Possible models

The experimental setup is open to a huge number of analyses and interpretations on the participants' part, inspired by past experience. As participants we can surmise that there's a connection between different trials, some sort of 'constant mechanism' at some level. Or we can surmise that there's no such connection, hence observation of past trials doesn't say anything about the next one. Or we can surmise that the next trial is influenced by the participant's own bar-height assignments. And many other hypotheses. We can also entertain all these hypotheses at the same time, and shift from one to another during the experiment. For example, if we suddenly wondered whether the computer program is actually using our bar distribution, we could suddenly move all probability to a slot at the edge and check if this seems to influence the next outcome (cf. participant 31).

The same goes for the choice of initial distribution. As participants we can say 'alright, there are 40 slots', and just give a uniform distributions to the 40 possibilities. Or we can consider the pyramidal mechanism of the game, which leads to a binomial distribution. Or we can consider that this is a computerized version of the game. The computer could simulate the physics of the actual game; but the image of the mechanism could also be just for show, the computer being programmed to distribute the outcomes according to a predetermined, completely arbitrary distribution. From this point of view we could again decide to assign a uniform distribution.

1.2 [Luca] Paradigms for judgement and assessment

The literature I've seen so far explains at length how the data presented to participants are generated, and is very succinct in explaining what was said to the participants before the experiment. The participant's inferences are then compared against models based on the pseudo-random algorithm that generated the data. Nassar et al.'s (2010) work is an example.

I think that we should use a different paradigm to describe the experiment and assess the participants' behaviours.

As I see it, the participants' inferential behaviours should be compared with that of a 'robot' that uses exact or approximate Bayesian or decision-theoretic rules, and that *starts from the same information that was given to the participants*. So it's really important that this information be explained at length, and whatever the participants were told should be reported verbatim.

I don't see the rationale of comparing a participant who doesn't know the data-generating algorithm, with a robot that does. Such a robot is modelling a different initial state of knowledge. What's important here, instead, is to model the inference, given the same initial knowledge.

A consequence of this point of view is that there isn't just one robot that can model the inferences. The information given to the participants is never enough to make numeric inferences and apply the probability calculus: it must always be augmented with additional assumptions, determined by each participant's previous life experiences. Different robots can thus be constructed: they use the same initial information as the participants, but each is augmented with different auxiliary assumptions. *The* ideal observer doesn't exist. There are several ideal observers.

Another consequence of this point of view is that the data-generating algorithm becomes slightly less important. The robot is constructed based on the exact information given the participants, and uses the same data given to the participants. The data-generating algorithm nowhere enters in the construction of the robot.

2 The Bayesian robot

2.1 A prosopopoeia of statistical models

A *statistical model* can be defined as a set of assumptions that allow us to make all possible numeric probabilistic inferences in the context of an experiment or of a sequence of observations. These assumptions correspond to – indeed they're the exact generalization of – axioms in propositional logic (Hailperin 1996; 2011; Jaynes 2003). Without axioms we can't derive theorems with the logical calculus; without initial probabilistic assumptions we can't derive or update probabilities with the probability calculus.

Consider this context: a set of M observations $1, 2, \ldots, M$, each of which can have N possible outcomes $\{1, \ldots, N\}$. Denote 'The outcome of the ith observation is d' by D^i_d. In this context, choosing a statistical model I is equivalent to choosing the numerical values of the joint distribution

$$\mathrm{P}(D^M_{d_M}, D^{M-1}_{d_{M-1}}, \ldots, D^2_{d_2}, D^1_{d_1} \mid I) \tag{1}$$

for all combinations of $d_1, \ldots, d_M \in \{1, \ldots, N\}$; or to choosing any other probabilistic assignments equivalent to this distribution. An equivalent alternative, for example, is to assign the distribution for the first observation, $\mathrm{P}(D^1_{d_1} \mid I)$, the N distributions for the second observation conditional on all possible first outcomes, $\mathrm{P}(D^2_{d_2} \mid D^1_{d_1}, I)$, and so on up to the N^{M-1} distributions for the Mth observation conditional on all possible previous outcomes, $\mathrm{P}(D^M_{d_M} \mid D^{M-1}_{d_{M-1}}, \ldots, D^1_{d_1}, I)$. The product of these distributions is the joint distribution above, and conversely each of these distributions can be obtained from the joint one above by marginalization and condizionalization.

Our assumptions can have different natures and motivations, for example from symmetry, mechanics, biology. As long as they lead to definite numerical values for the distribution above, they constitute a statistical model.

The term 'statistical model' is not universally used this way, though.[1]

[1] 'The conflict regarding the existence of a or the true model, model uncertainty, identification, selection, misspecification, etc. can be resolved if a clear definition of a model is stated. Models have different meanings in different fields. Once a young woman was interviewed for a data handling job in Unilever. She was puzzled by one of the interviewers who asked her whether she had done any modelling. It was found later that to her the question meant whether she had posed for photographs!' J. R. M. Ameen (Copas et al. 1995 p. 453).

It often has a more general meaning – what we'd call 'classes of statistical models' in our terminology.

To avoid using this ambiguous term, we'll figuratively speak of a *Bayesian robot* instead of a model. Such a robot uses the three laws of probability to make numeric probabilistic inferences and update them in view of new observations. But to do so it first needs to be fully programmed with a complete set of probabilistic assumptions, in the form of the joint distribution (1) or in another equivalent way. *A 'Bayesian robot' is thus equivalent to a statistical model.*

In the context of the experiment described above, the set of possible robots is a simplex of $N^M - 1$ dimensions, since this is the set of all possible distributions (1) for M joint outcomes, each having N possible values.

In this study, the term 'inference' will be synonymous with 'assignment of a probability (distribution)', just to keep sentences short. Our statistical terminology and notation follow ISO standards (ISO 2009; 2006), except for the use of a comma to denote logical conjunction '$\wedge$', for simplicity.

2.2 Participants' statistical models in the Plinko experiment

In the Plinko experiments (Filipowicz et al. 2014; 2016) the set of possible Bayesian robots is a simplex of $40^{200} - 1 \approx 3 \times 10^{320}$ dimensions. The choice of a specific robot, or at least the restriction of the set of robots to use, is determined by the verbal information given before the Plinko task. In the Plinko experiments the participants aren't given any particular information about the source or sources of the outcomes ✠ Luca: as far as I've understood . Thus no restriction is placed on the choice of robot. We could maybe appeal to 'common sense' assumptions, but who's to say what's common sense? Each participant could have widely different past experiences; what's common sense to us may not be to them and vice versa.

The distributions sequentially constructed by a participant represent the M conditional distributions

$$\mathrm{P}(D^m_{d_m} \mid D^{m-1}_{d_{m-1}}, \dots, D^1_{d_1}, \text{participant}), \qquad m = 1, \dots, M,$$
$$\text{conditional on the observed } d_1, \dots, d_{m-1} \text{ only.} \quad (2)$$

These distributions are not sufficient to define the joint distribution (1): they only give us $39 \times 200 = 7\,800$ out of the 3×10^{320} required numbers. The participant's statistical model is thus largely unknown.

The participant's inferences can indeed come from several possible statistical models: there are several Bayesian robots that would make the exact same inferences as the participant. The simplest such robot, call it I_p, is the one for which

$$\mathrm{P}(D^m_{d_m} \mid D^{m-1}_{d_{m-1}}, \dots, D^1_{d_1}, I_\mathrm{p}) = \mathrm{P}(D^m_{d_m} \mid I_\mathrm{p}) = \text{participant's} \tag{3a}$$

which yields a normalized joint distribution

$$\mathrm{P}(D^M_{d_M}, D^{M-1}_{d_{M-1}}, \dots, D^2_{d_2}, D^1_{d_1} \mid I_\mathrm{p}) = \mathrm{P}(D^M_{d_M} \mid I_\mathrm{p}) \times \cdots \times \mathrm{P}(D^1_{d_1} \mid I_\mathrm{p}) \qquad \text{for all } d_1, \dots, d_M, \tag{3b}$$

which completely defines the robot's inferences. This robot assumes the probabilities for all observations to be independent of one another, and they happen to be equal to the participant's. One may argue that this is an 'unreasonable' robot, and that the participants do not assume independence of the probabilities. But *there's no way to prove this* from the verbal information given to the participants and the set-up of the Plinko experiments (Filipowicz et al. 2014; 2016).

Since the initial information given to the participants places no restrictions is placed on the set of possible statistical models, the model (3) is legitimate. Each participant is thus making inferences in a perfectly Bayesian way.

If the participants were given more detailed information, for example 'All outcomes are generated by the same computer algorithm', or 'Two computer algorithms generate the outcomes; one will succeed the other after an unknown number of observations', or 'The generating algorithm does not use knowledge of previous outcomes', then this information would restrict the set of allowed Bayesian robots, and we could check whether each participant was behaving in a Bayesian way or not (assuming no misunderstanding of the initial information occurred).

Another way to get a glimpse of a participant's statistical model is to let him or her do the full Plinko experiment several times, asking to start with the same initial distribution and making it very clear that the same set-up is used in all repetitions of the experiment.

In the next two sections we consider two sets of robots programmed with two different initial assumptions.

3 First study: the Johnson-Dirichlet robots

3.1 A first Bayesian robot

We now build a set of robots programmed with particular initial assumptions, denoted I_1, leading to a specific joint distribution (1).

We first require that joint distribution to satisfy this property:

- The joint probability distribution (1) for *any* number of observations is invariant with respect to exchanges of their order.

This property corresponds to either of these assumptions:

(*a*) the order of the observations is irrelevant for making inferences;

(*b*) only the relative frequencies of known observations are relevant for making inferences; any additional data about known observations is irrelevant and can be discarded.

The equivalence of property (4) with assumption (*a*) is evident. Less so the equivalence with assumption (*b*); it will be shown in § 3.2.

If the participants were initially given either of these pieces of information, the joint distribution formed by their sequence of distributions would have to satisfy property (4).

Property (4) above is called *infinite exchangeability*. This notion was introduced by de Finetti (1930; 1937; Heath et al. 1976). Dawid (2013) gives a non-technical, insightful introduction to it. It is studied in detail in Bernardo & Smith (2000 § 4.2). A theorem by de Finetti shows that infinite exchangeability implies that the probability distribution (1) must have this mathematical form, for any m:

$$\mathrm{P}(D^1_{d_1}, D^2_{d_2}, \dots, D^m_{d_m} \mid I_1) = \int_\Delta \left(\prod_{i=1}^{m} q_{d_i} \right) \mathrm{p}(\boldsymbol{q} \mid I_1)\, \mathrm{d}\boldsymbol{q}, \tag{4}$$

where $\boldsymbol{q}$ is a normalized N-tuple of positive numbers:

$$\Delta \coloneqq \{\boldsymbol{q} \in \mathbf{R}^N \mid q_d \geqslant 0, \textstyle\sum_{d=1}^{N} q_d = 1\}, \tag{5}$$

and $\mathrm{p}(\boldsymbol{q} \mid I_1)\, \mathrm{d}\boldsymbol{q}$ is a normalized density function, called *prior parameter density*.

The N-tuple $\boldsymbol{q}$ can be interpreted as the relative long-run frequencies of the possible outcomes[2], and $\mathrm{p}(\boldsymbol{q} \mid I_1)\, \mathrm{d}\boldsymbol{q}$ as their probability density.

[2] ‘But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead.’ (Keynes 2013 § 3.I, p. 65)

From this point of view it is as if the robot first assumes to know the long-run frequencies of the different N outcomes, and assumes that the outcome sequences showing these frequencies are all equally probable. The probability for each outcome in the forthcoming observations is thus proportional to its frequency:

$$\mathrm{p}(D^1_{d_1}, D^2_{d_2}, \dots, D^m_{d_m} \mid \boldsymbol{q}, I_1) = \prod_{i=1}^{m} q_{d_i}. \tag{6}$$

Then, being uncertain about the long-run frequencies, the robot assigns to them the density $\mathrm{p}(\boldsymbol{q} \mid I_1)\,\mathrm{d}\boldsymbol{q}$. The combined uncertainty about the observations is then eq. (4) by the law of total probability.

The assumption of infinite exchangeability is not enough to have a fully programmed robot, because it doesn't determine the parameter density $\mathrm{p}(\boldsymbol{q} \mid I_1)\,\mathrm{d}\boldsymbol{q}$. Once this is chosen – for example, the constant density $(N-1)!\,\mathrm{d}\boldsymbol{q}$, called Bayes's (1763 Scholium) or Laplace's (1819 p. xvii) prior – the robot is fully programmed. We consider a specific density in § 3.4, and consider it as part of the assumptions I_1.

In fact, the robot resulting from the assumptions I_1 is more powerful than necessary, because it can make inferences about *any* number of unknown observations, future or past ones, beyond the M considered in our context.

An explicit example of formula (4) with $N = 40$ is

$$\mathrm{P}(D^1_{37}, D^2_{6}, D^3_{25}, D^4_{37} D^5_{19} \mid I_1) = \int_{\Delta} q_6\, q_{19}\, q_{25}\, {q_{37}}^2\, \mathrm{p}(\boldsymbol{q} \mid I_1)\,\mathrm{d}\boldsymbol{q}. \tag{7}$$

In the following we omit the integration domain Δ.

3.2 Probability updates from observations

Using Bayes's theorem the robot can calculate from (4) the probability distribution for the mth observation, having observed outcomes $d_1, \dots, d_{m-1}$,

$$\mathrm{P}(D^m_{d_m} \mid D^{m-1}_{d_{m-1}}, \dots, D^2_{d_2}, D^1_{d_1}, I_1), \tag{8}$$

obtaining

$$\mathrm{P}(D^m_{d_m} \mid D^{m-1}_{d_{m-1}}, \dots, D^2_{d_2}, D^1_{d_1}, I_1) = \int q_{d_m}\, \mathrm{p}(\boldsymbol{q} \mid D^1_{d_1}, \dots, D^{m-1}_{d_{m-1}}, I_1)\,\mathrm{d}\boldsymbol{q}, \tag{9a}$$

with

$$\mathrm{p}(\boldsymbol{q}|\, D^1_{d_1}, \dots, D^{m-1}_{d_{m-1}}, I_1) = \frac{\left(\prod_{i=1}^{m-1} q_{d_i}\right) \mathrm{p}(\boldsymbol{q}|\, I_1)}{\int \left(\prod_{i=1}^{m-1} q'_{d_i}\right) \mathrm{p}(\boldsymbol{q}'|\, I_1)\, \mathrm{d}\boldsymbol{q}'}. \tag{9b}$$

Once the density $\mathrm{p}(\boldsymbol{q}|\, I_1)\, \mathrm{d}\boldsymbol{q}$ is chosen, the robot can thus perform the Plinko task just like all others participants.

For our numeric example (7) this is

$$\mathrm{P}(D^5_{19}|\, D^1_{37}, D^2_6, D^3_{25}, D^4_{37}, I_1) = \int q_{19}\; \mathrm{p}(\boldsymbol{q}|\, D^1_{37}, D^2_6, D^3_{25}, D^4_{37}, I_1)\, \mathrm{d}\boldsymbol{q}, \tag{10a}$$

$$\mathrm{p}(\boldsymbol{q}|\, D^1_{37}, D^2_6, D^3_{25}, D^4_{37}, I_1) = \frac{q_6\, q_{25}\, {q_{37}}^2\; \mathrm{p}(\boldsymbol{q}|\, I_1)}{\int q'_6\, q'_{25}\, {q'_{37}}^2\; \mathrm{p}(\boldsymbol{q}'|\, I_1)\, \mathrm{d}\boldsymbol{q}'}. \tag{10b}$$

Formulae (9b) and (10) show that the updated conditional distribution depends only on the number of times each outcome appeared in the known observations, that is, on their frequencies. For example, in (10) outcomes 6 and 25 appeared once each (exponent of q_d is 1), outcome 37 appeared twice (exponent is 2), and the others never appeared (exponent is 0). The order of their appearance does not enter the formula. This shows the equivalence of infinite exchangeability with the assumption *(b)* of the previous section. The reverse equivalence can be proved using a series of theorems by Koopman, Pitman, Lauritzen ✠ add refs .

3.3 General remarks on the Johnson-Dirichlet robot's behaviour

The infinite-exchangeability formulae (4) and (9) lead to peculiar features of the robot's inferences and of their updates:

- One way to interpret the robot's inferences is this: the robot assumes that there are 'constant conditions' in all trials; loosely speaking, a 'constant mechanism'. The outcome of each observation is therefore not influenced by the outcomes of other observations. Another interpretation is this: the robot does not keep track of the order of the observations, because this information is irrelevant; but keeps track of the outcome frequencies. Trends are therefore invisible to the robot.
- As data D accumulate, the updated density $\mathrm{p}(\boldsymbol{q}|\, D, I_1)\, \mathrm{d}\boldsymbol{q}$ becomes more and more peaked at the N-tuple of observed frequencies. The

robot's probabilities for the next outcome will therefore approach the frequencies hitherto observed. This approach happens independently of the form of the density $\mathrm{p}(\boldsymbol{q}|\, I_1)\,\mathrm{d}\boldsymbol{q}$ – unless the latter is zero in peculiar regions of the integration domain – but this density determines the celerity of the approach. A density heavily peaked on a frequency $\boldsymbol{q}'$ will require a lot of data to move the predictions to a very different frequency $\boldsymbol{q}''$.

- Suppose that robot receives a long sequence of observations concentrated around frequencies $\boldsymbol{q}'$ – say, a very long sequence of 3s in a row – and then another long sequence concentrated around different frequencies $\boldsymbol{q}''$ – say, suddenly only 5s appear. After the shift, the robot's probabilities will eventually become peaked around the new frequencies, but the shift in the peaks will take a larger number of observations around the new frequencies than the number around the old frequencies.

3.4 Prior parameter density

The shape of the prior parameter density heavily determines the first inferences. Note that the Plinko experiment tells us each participant's probabilities for the first observation,

$$\mathrm{p}(D_d^1|\,\text{participant}) \equiv \int q_d\ \mathrm{p}(\boldsymbol{q}|\,\text{participant})\,\mathrm{d}\boldsymbol{q}, \tag{11}$$

but this does not determine the prior parameter density $\mathrm{p}(\boldsymbol{q}|\,\text{participant})\,\mathrm{d}\boldsymbol{q}$.

In this first study we consider a *Johnson-Dirichlet* prior density (references below). This density is proportional to a monomial in (q_i):

$$\mathrm{p}(\boldsymbol{q}|\, I_1) = \frac{\Gamma(\Lambda)}{\prod_i \Gamma(\Lambda\,\nu_i)} \prod_{i=1}^{N} {q_i}^{\Lambda\,\nu_i - 1}, \qquad \Lambda > 0,\, \nu \in \Delta. \tag{12}$$

The N independent parameters $(\Lambda\,,\,\nu)$ need to be specified by additional assumptions.

Such prior density is determined by the assumption that that the frequencies of other outcomes are irrelevant for predicting a particular

one. If $\boldsymbol{f} \coloneqq (f_1, \dots, f_N)$ are the observed outcomes' relative frequencies, this assumption can be mathematically expressed as

$$\mathrm{P}(D_d^{m+1} | \, \boldsymbol{f}, N, I_1) = \mathrm{P}(D_d^{m+1} | \, f_d, N, I_1) \qquad \text{for all } d \in \{1, \dots, N\}. \quad (13)$$

It is called 'sufficientness' (Johnson 1924; 1932; Good 1965 ch. 4; Zabell 1982; Jaynes 1996). We consider this assumption and the specification of the parameters above to be contained in I_1.

The density (12) is a conjugate prior (DeGroot 2004 ch. 9; Diaconis et al. 1979). It has two convenient properties: it updates to a density of the same mathematical form, and leads to probabilities that can be analytically calculated using the formula

$$\int_\Delta \prod_{i=1}^{N} {q_i}^{x_i - 1} \, \mathrm{d}\boldsymbol{q} = \frac{\prod_i \Gamma(x_i)}{\Gamma\bigl(\sum_i x_i\bigr)}. \quad (14)$$

In fact, inserting this prior in the infinite-exchangeability formula (9) we find the joint distribution (1), for any m:

$$\mathrm{p}(D_{d_1}^1, \dots, D_{d_m}^m | \, I_1) = \frac{\Gamma(\lambda)}{\Gamma(\lambda + m)} \prod_{d=1}^{N} \frac{\Gamma(\lambda \nu_d + m f_d)}{\Gamma(\lambda \nu_d)}, \quad (15)$$

where f_d is the frequency of outcome d in the m observations.

In particular, the initial distribution is simply

$$\mathrm{P}(D_d^1 | \, I_1) = \nu_d, \quad (16)$$

for this reason we call this parameter the *initial distribution* of the robot. This formula says that the Johnson-Dirichlet prior can produce any possible initial probability distribution, by appropriately choosing $\boldsymbol{\nu}$. This freedom will be used in the next section to mimic the participants' behaviours.

We call the parameter λ *stubbornness* of the robot; here's the reason.

From eqs (9b), (12), and (14) we can calculate the parameter density updated after the observations $1, \dots, m-1$:

$$\mathrm{p}(\boldsymbol{q} | \, D_{d_1}^1, \dots, D_{d_{m-1}}^{m-1}, I_1) = \frac{\Gamma(\lambda')}{\prod_i \Gamma(\lambda' \nu_i')} \prod_{i=1}^{N} {q_i}^{\lambda' \nu_i' - 1}$$

$$\text{with} \quad \lambda' = \lambda + m - 1, \quad \boldsymbol{\nu}' = \frac{\lambda \boldsymbol{\nu} + (m-1)\boldsymbol{f}}{\lambda + m - 1} \quad (17)$$

where $\boldsymbol{f}$ are the relative frequencies of the outcomes in the $m-1$ observations. This formula, together with (9) and (14), yields the robot's inference for the mth observation conditional on the previous ones:

$$\mathrm{P}(D^m_d \mid D^{m-1}_{d_{m-1}}, \dots, D^1_{d_1}, I_1) = \frac{\Lambda\, \nu_d + (m-1)\, f_d}{\Lambda + m - 1}. \tag{18}$$

The conditional probability for the mth observation has an interesting interpretation: it is equal to a weighted sum of the initial distribution (16) and the relative frequency distribution $\boldsymbol{f}$ in the $m-1$ previous observations. The weight of the former is the parameter Λ, and of the latter the number $m-1$ of observations. If the parameter Λ is very large with respect to the number of observations, then the successive conditional distributions remain approximately equal to the initial one $\boldsymbol{\nu}$. If Λ is very small, the robot quickly changes its initial distribution into the distribution of observed frequencies instead. Hence Λ quantifies how 'stubborn' the robot is in keeping its initial belief. The robot behaves as if it had already made Λ observations with outcome frequencies $\boldsymbol{\nu}$. In general it takes Λ observations to make the robot appreciably change its initial belief.

Owing to the special mathematical form of update formulae above, it's possible to completely characterize the 'beliefs' of our robot after each observation with N independent parameters, whose values are determined by the values before that observation and the outcome observed. Denote the parameter values before the ith observation by $\bigl(\Lambda^{(i-1)}, \boldsymbol{\nu}^{(i-1)}\bigr)$. If the outcome of that observation is d_i, then the values after the observation are

$$\begin{aligned} \Lambda^{(i)} &\coloneqq \Lambda^{(i-1)} + 1, \\ \nu_d^{(i)} &\coloneqq \frac{\Lambda^{(i-1)}}{\Lambda^{(i-1)}+1}\, \nu_d^{(i-1)} + \begin{cases} \dfrac{1}{\Lambda^{(i-1)}+1} & \text{for } d = d_i, \\ 0 & \text{for } d \neq d_i. \end{cases} \\ &\text{with} \qquad \Lambda^{(0)} \coloneqq \Lambda, \quad \boldsymbol{\nu}^{(0)} \coloneqq \boldsymbol{\nu}. \end{aligned} \tag{19}$$

From (19)(17) one can show that $\boldsymbol{\nu}^{(i)}$ is also the conditional probability for the $(i+1)$th observation, given the previous observations.

This update corresponds to a participant's raising, on the screen, the prediction bar under slot d_i by a particular amount, leaving the others untouched; or to lowering the prediction bar for all other slots in the same proportion, leaving the d_i untouched.

3.5 Examples: participant vs robot

The results of the previous sections show that we can program our robot to make the same initial inference as a participant, just by setting its parameters $\boldsymbol{\nu}^{(0)} \coloneqq \boldsymbol{\nu}$ equal to the participant's initial distribution. We are still free to choose $\Lambda^{(0)} \coloneqq \Lambda$, but it's generally impossible to find a value that will lead to the participant's following sequence of distributions; already the second will generally be different. When this matching is possible it means that the participant is behaving according to the model I_1.

Let's choose a participant and set the robot's initial distribution equal to his or hers. We can examine how the robot make the following inferences, based on the same sequence of outcomes observed by the participant, for several values of the stubbornness Λ. Figure 1 shows the means and standard deviations of the sequence such distributions, for participant 12 and a robot with stubbornness $\Lambda = 0.1$. This low value makes the robot give great consideration to the first outcomes, as the initial high variability of the robot's means shows. The algorithm generating the Plinko outcomes had a change in standard deviation after the 100th outcome, shifting to a narrower distribution. The robot adapted to this change very slowly, because it stubbornness at that point was $\Lambda^{(100)} \approx 100$.

Figure 2 is analogous to fig. 1 but for a robot with stubbornness $\Lambda = 50$. This robot is even more slow to adapt to the narrowing in the standard deviation of the generated outcomes.

Figures 3 and 4 show the same for participant 30. The change in standard deviation was from narrow to large in this case.

The robot with low stubbornness seems to adapt to the widening of the outcome outputs faster than it had for the narrowing of the previous case: the change in the slope of the robot's standard-deviation curve seems steeper in fig. 3 than in 1.

If we look at the sequence of outcomes of figs 1 or 3, we perceive that something changed around the 100 observation. If we could plot these outcomes while they are generated, we would likely notice the change by around the 25th observation. Our robot, however, can't detect this change for the reasons explained in § 3.3; any outcomes from narrow or wide generating processes are mixed in the robot's memory.

Only non-exchangeable or hierarchic statistical models, like the one we develop in § 4, can exhibit a sort of memory for the order and be

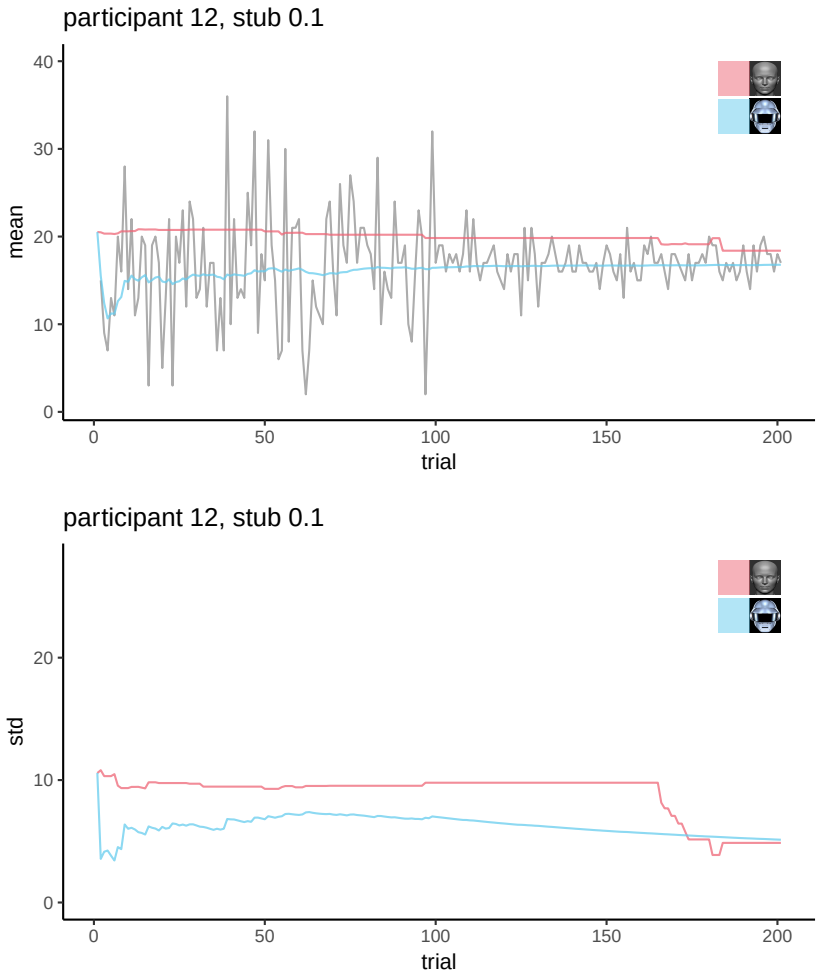

Figure 1 Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubbornness $\Lambda = 0.1$

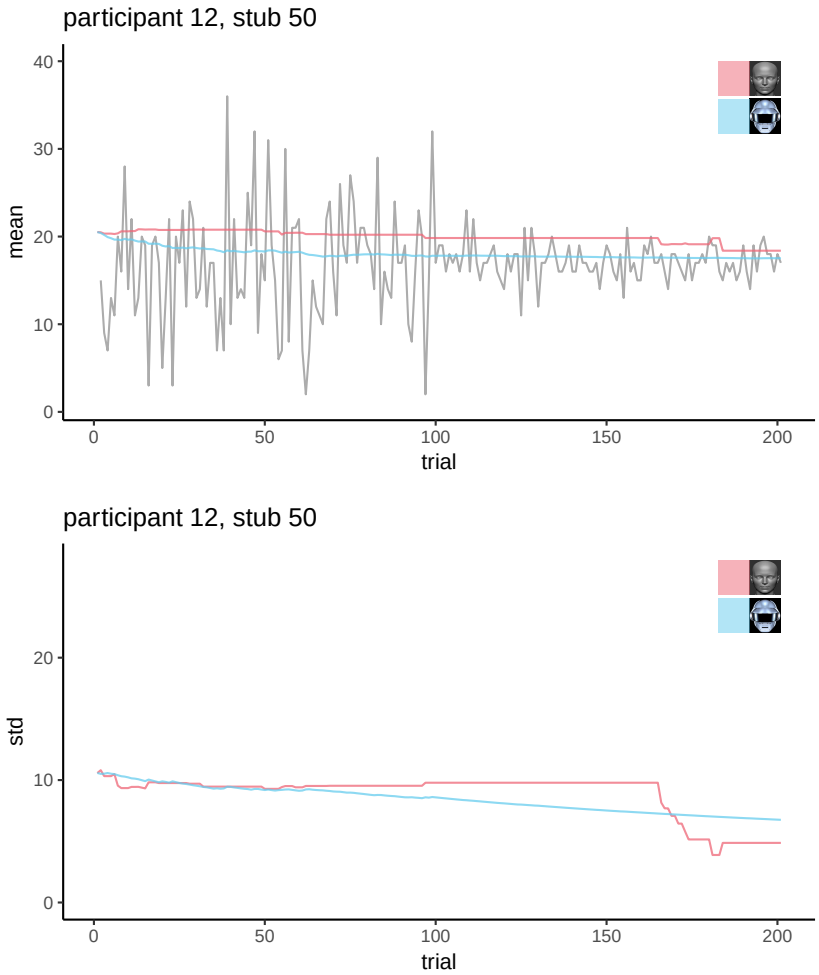

Figure 2 Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubbornness $\Lambda = 50$

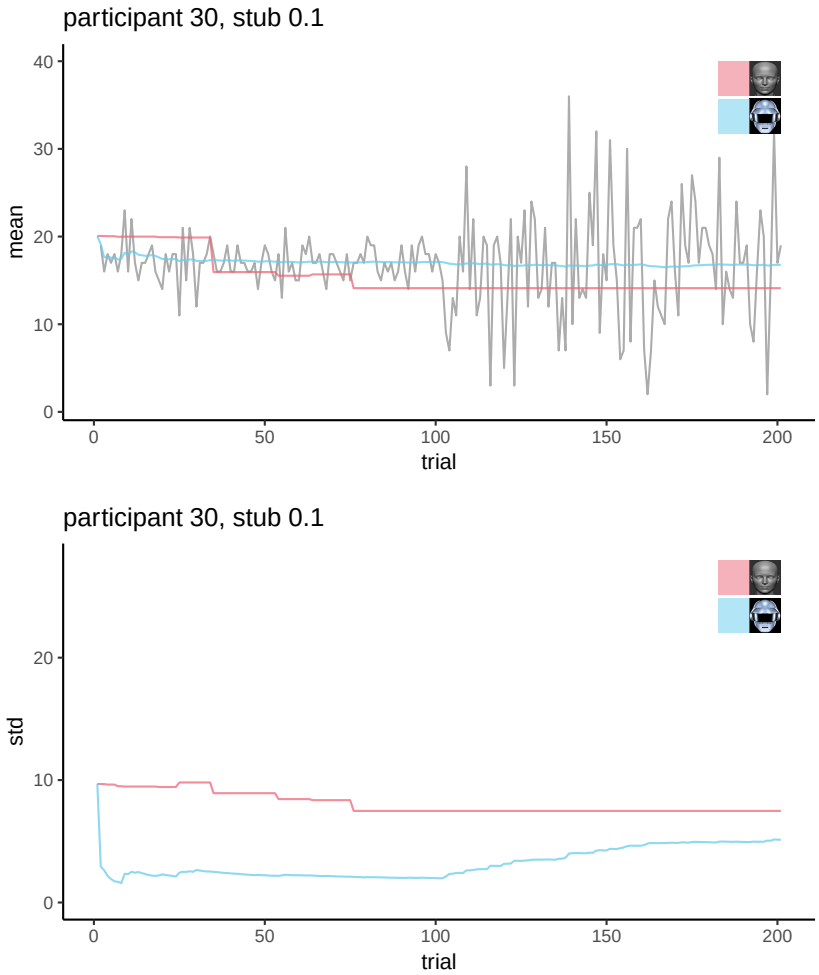

Figure 3 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubbornness $\Lambda = 0.1$

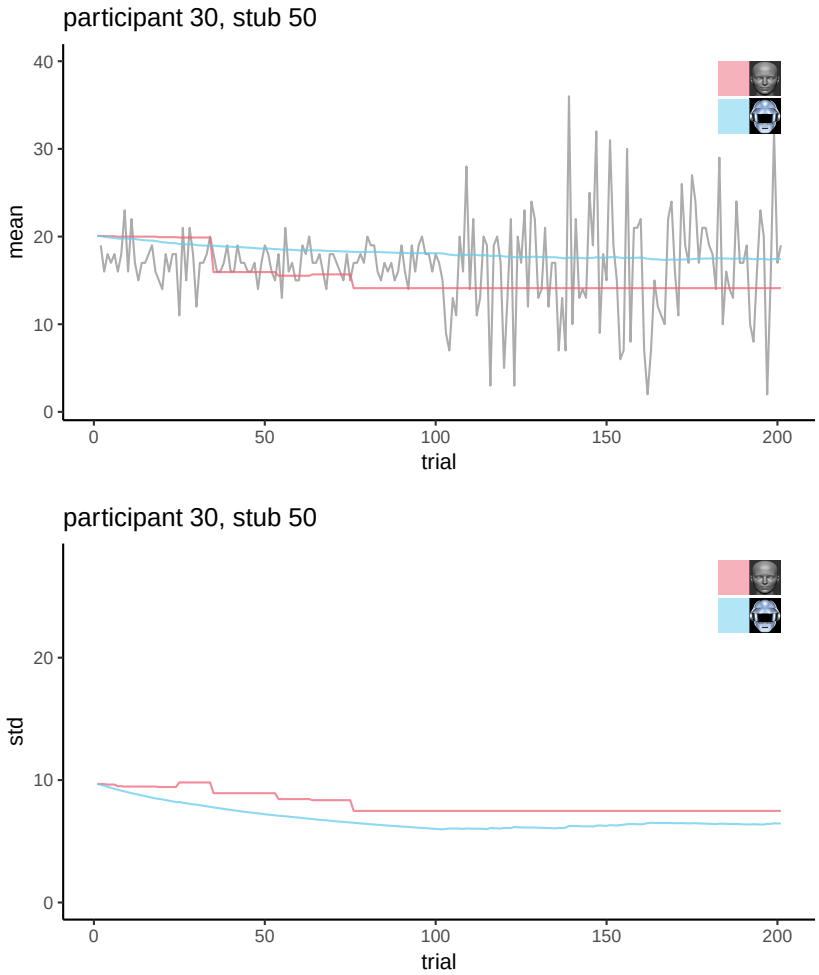

Figure 4 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubbornness $\Lambda = 50$

capable of believing that the underlying 'mechanism' has changed.

Some conclusions can be drawn from the properties of our model and from the examples:

- Participants who have great inertia against updating their predictions in view of the observations are *not* necessarily behaving at variance with the probability calculus. The latter says that they can be as stubborn as they please: larger Λ. If we judge such inertia as irrational, our judgement cannot be based on such a simple model; possibly it's based on a hierarchic model where Λ is given a probability that depends on past experiences.
- The slots have a specific physical order, and from the way the ball falls into them it seems reasonable to assume that updates to the probability for one slot should affect those for nearby slots. The Johnson-Dirichlet model does not take this into account.
- Infinitely exchangeable priors are incapable of quickly adapting to changes in the empirical statistics of the outcomes.

3.6 The robot's surprise

A robot with low stubbornness quickly adapts its predictions to the observed outcomes, but as the outcomes accumulate its stubbornness increases. If there is a late change in the generation of the outcomes, after the 100th observation for example, the robot will adapt its predictions more slowly.

It is interesting to ask: is this prediction adaptation the same for a change from a narrow to a wide distribution, as for a change from a wide to a narrow one? or is there a difference in the adaptation speed?

The answer depends on how we measure such speed. One way could be this: starting from the observation in which the change occurs, we let a second robot with low stubbornness observe the new outcomes and make predictions, starting from frequency parameters equal to those reached by the first robot. The second robot will quickly adapt to the new observed outcomes, and we can use it as a touchstone for the first robot's adaptation speed. The predictions of the second robot can also be interpreted as if we had reset the stubbornness of the first robot to a low value.

The results of this comparison are shown in fig. 5. Looking at the standard deviations of the distributions it seems that a robot adapts

more slowly in going from a wide to a narrow distribution than vice versa. This is true looking at the final relative entropy of the first robot relative to the second: 1.01 wide-to-narrow vs 0.260 narrow-to-wide;

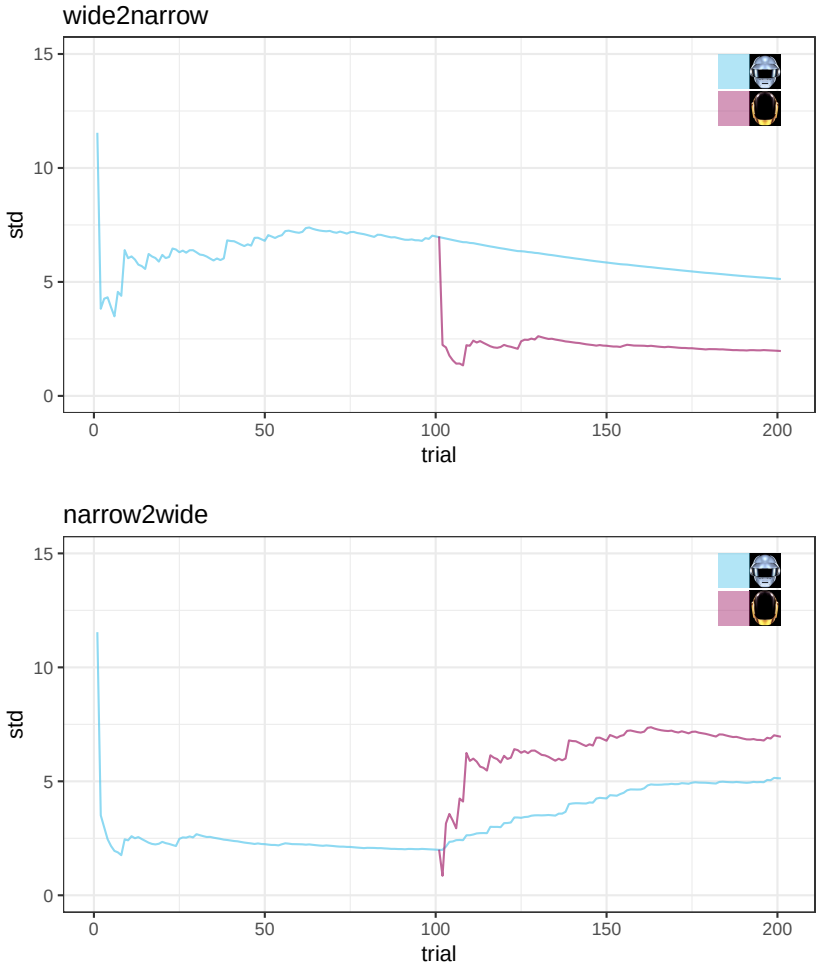

Figure 5 Adaptation speed of the standard deviation for a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change. The final relative entropy for the first robot relative to the second is 1.01 in the wide-to-narrow case, vs 0.260 in the narrow-to-wide; exchanging the distributions: 0.283 vs 0.211. The final normalized overlap is 0.949 for wide-to-narrow vs 0.823 for narrow-to-wide

same if we exchange the distributions: 0.283 vs 0.211. The overlap seems to say the opposite: 0.106 wide-to-narrow vs 0.0512 narrow-to-wide; but the overlap is heavily influenced by the narrowness of the overlapping distributions, so it may not be a reliable measure in this case. If we use the normalized overlap (corresponding to the cosine of the angle between the distribution vectors) we find 0.949 vs 0.823, which agree with the first three measures.

Figure 6 shows that this phenomenon is even more striking if we average the sequences of standard deviations over 100 repetitions of such experiments.

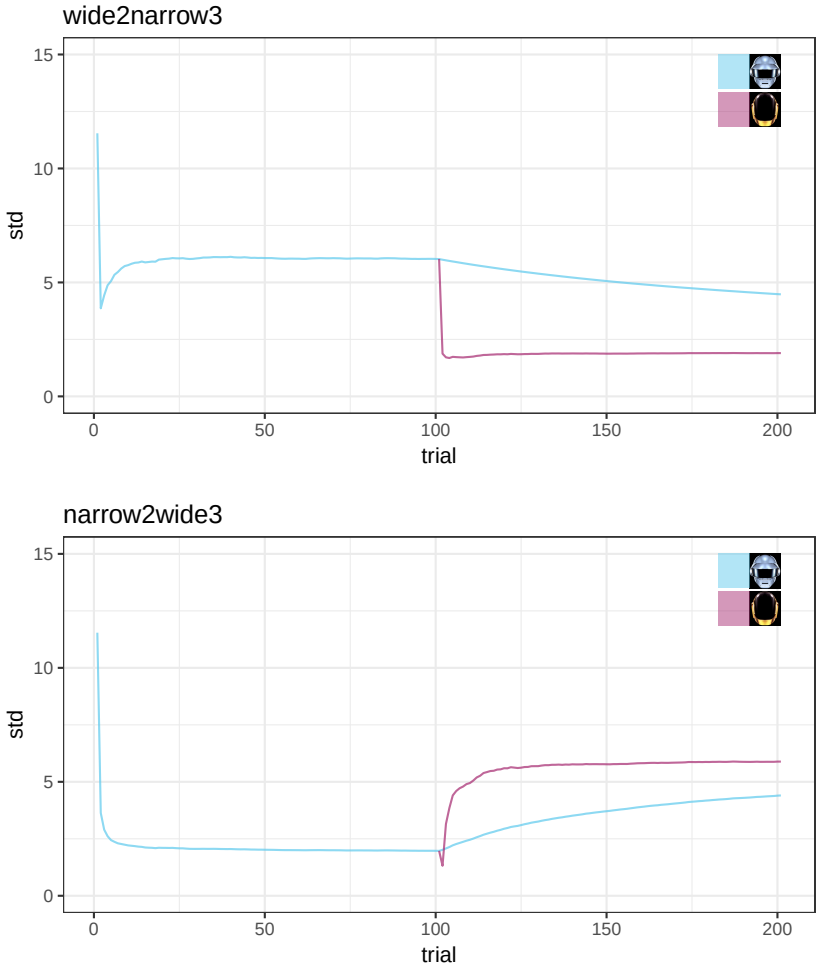

Figure 6 Adaptation speed of the standard deviation for a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change, averaged over 100 experiments. The normals have same mean 17 and standard deviations 6 and 1.9.

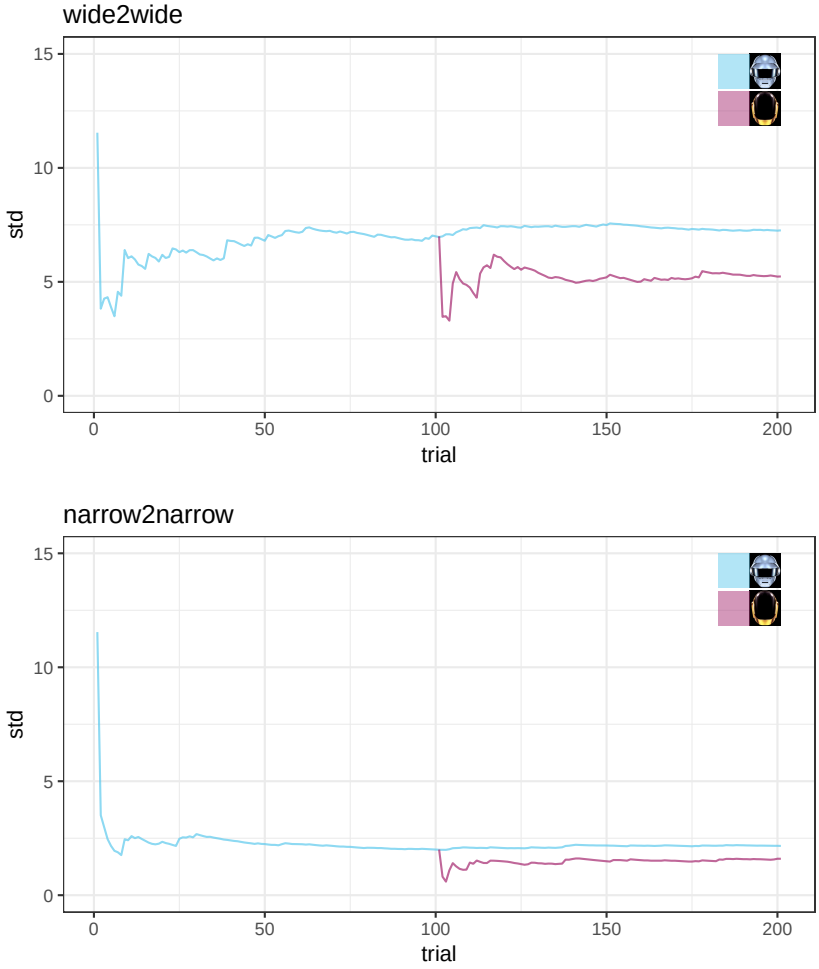

Figure 7 Adaptation speed of a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change.

4 Second study: Johnson-Dirichlet robots with change-points

4.1 The robot looks for changepoints

The robot of the previous study remembers past outcomes but not their order. Said otherwise, it assumes that past frequencies are all that matters for its inference. We now consider a new robot that takes the outcome order into account in a simple way.

This new robot, denoted I_2, is based on this assumption:

- the sequence of observations is divided into subsequences of contiguous observations. These subsequences can have different lengths. Within each subsequence, the order of the observations doesn't matter, and the Johnson-Dirichlet assumptions apply locally.

This robot therefore uses an infinitely exchangeable model with a Johnson-Dirichlet parameter density for each subsequence, as in the previous study. Only observations within a subsequence contribute to the update of the distributions for the other observations in that subsequence. Observations from other subsequences are ignored. As one subsequence ends, the robot's initial-distribution parameter $\boldsymbol{\nu}$ and stubbornness Λ are reset.

The robot doesn't know the changepoints in advance, however. Before each observation, the robot first infers whether the observation belongs to a new subsequence or it continues the previous one. In the first case, it then infers the next outcome using a fresh Johnson-Dirichlet model. In the second case, it infers the next outcome updating the parameters of Johnson Dirichlet model used thus far. The final inference is a combination of these two according to the law of total probability.

To translate these assumptions into probability distributions it's convenient to introduce some propositions, denoted R_r^m, defined this way:

$$\begin{aligned} R_r^m &\coloneqq \text{'Only the previous } r \text{ outcomes are relevant for the } m\text{th one'} \\ &\equiv \text{'Observations } \{m-r, m-r+1, \dots, m\} \text{ form a subsequence'} \\ &\qquad\qquad \text{with } 0 \leqslant r \leqslant m-1. \end{aligned} \tag{20}$$

The equivalence of these two propositions should be clear from the discussion above.

The proposition R_0^m, in particular, states that *no* previous observations are relevant for inferring the mth one; in other words, it states that this observation is the start of a new subsequence.

For fixed m, the propositions R_r^m, $0 \leqslant r \leqslant m-1$, are obviously exhaustive (their probabilities sum to unity) and mutually exclusive. The assumptions I_2 made thus far imply that some propositions for different m are also exclusive:

$$\mathrm{P}(R_{r_m}^m, R_{r_{m-1}}^{m-1} \mid I_2) = 0 \quad \text{if } r_m \neq 0 \text{ and } r_m \neq r_{m-1} + 1, \tag{21}$$

and therefore

$$\mathrm{P}(R_0^m \mid R_s^{m-1}, I_2) + \mathrm{P}(R_{s+1}^m \mid R_s^{m-1}, I_2) = 1. \tag{22}$$

The reason is that the mth outcome either starts a new subsequence, hence $r_m = 0$; or continues the current subsequence, which according to $R_{r_{m-1}}^{m-1}$ contains the $(m-1)$th outcome and the previous r_{m-1} ones, and therefore $r_m = r_{m+1} + 1$. All other cases are impossible. Note that these other cases might be possible under assumptions different from I_2; they would represent a robot with a sort of moving-window memory.

With these propositions about the changepoints, the reset rules described above correspond to the following equalities, for all m:

$$\mathrm{P}(D_{d_m}^m \mid R_r^m, D_{d_{m-1}}^{m-1}, \dots, D_{d_1}^1, I_2) = \begin{cases} \mathrm{P}(D_{d_m}^m \mid R_r^m, D_{d_{m-1}}^{m-1}, \dots, D_{d_{m-r}}^{m-r}, I_2) & \text{for } 1 \leqslant r \leqslant m-1, \\ \mathrm{P}(D_{d_m}^m \mid R_r^m, I_2) & \text{for } r = 0, \end{cases} \tag{23}$$

Where the probabilities on the right side are given by the Johnson-Dirichlet formulae (18) or (19).

The formulae above lead to a conditional distribution for the mth outcome given the previous ones only if the robot has been programmed with the probabilities

$$\mathrm{P}(R_{r_m}^m \mid D_{d_{m-1}}^{m-1}, \dots, D_{d_1}^1, I_2), \tag{24}$$

which in turn can be obtained by marginalization and condizionalization from the probabilities

$$\mathrm{P}(R_{r_m}^m \mid D_{d_{m-1}}^{m-1}, R_{r_{m-1}}^{m-1}, \dots, R_{r_2}^2, D_{d_1}^1, I_2) \tag{25}$$

for all legitimate m, r_i, and outcomes d_i. Additional assumptions are necessary to determine the latter probabilities. For example, the particular

order of all past outcomes, the lengths of the previous subsequences, or even the robot's previous distributions, and other factors external to the experiment might all be relevant. The robot is programmed with this assumption:

- only two quantities are relevant to infer whether the next observation starts a new subsequence: the total number $m-1$ of observations made so far, and the current length s of the present subsequence.

In formulae this assumption is equivalent to a Markov property:

$$\mathrm{P}(R^m_{r_m} \mid D^{m-1}_{d_{m-1}}, R^{m-1}_{r_{m-1}}, \dots, R^2_{r_2}, D^1_{d_1}, I_2) = \mathrm{P}(R^m_{r_m} \mid R^{m-1}_{r_{m-1}}, I_2). \tag{26}$$

Recalling, eq. (22), that only the values $r_m = 0$ and $r_m = r_{m-1} + 1$ are possible, the conditional probabilities above are summarized in the *changepoint function*

$$h(s,m) \coloneqq \mathrm{P}(R^m_0 \mid R^{m-1}_s, I_2), \qquad m \in \{2,3,\dots\}, \quad s \in \{1,\dots,m-1\}. \tag{27}$$

It is the probability that the mth observation starts a new subsequence, given that the previous subsequence had length s, including the $(m-1)$th outcome (that's the reason for $s-1$ in the definition). Note that from (21)

$$\mathrm{P}(R^m_s \mid R^{m-1}_{s-1}, I_2) = 1 - h(s,m), \qquad s \neq 0. \tag{28}$$

A little counting shows that the set of the possible changepoint functions, for an experiment with M observations, is $[0,1]^{(M^2-M)/2}$.

Once we specify:

- the stubbornness Λ and initial distribution $\boldsymbol{\nu}$ to be used at each reset,
- the changepoint function $h(s,m)$,

our new robot I_2 is completely programmed and can face the Plinko task. Note that these specifications determine the joint distribution (1), through the assumptions made so far.

4.2 Online belief update

The beliefs of the Johnson-Dirichlet robot of § 3 during the Plinko task were fully summarized by N independent parameters, recursively updated online after each observation.

An analogous online update is possible for this robot, with some differences: first, the number of parameters required increases linearly

with the observations; second, the updated parameter values depend on the last outcome, on the parameter values before the last observation, *and* on some of the parameter values before the last but one observation. The online-update algorithm we now develop follows closely Adams & MacKay's (2007) but departs from it in some important respects.

We first write probability for the mth observation given the previous outcomes can be written

$$\mathrm{P}(D^m_{d_m} \mid D^{m-1}_{d_{m-1}}, \ldots, D^1_{d_1}, I_2) = \frac{\mathrm{P}(D^m_{d_m}, D^{m-1}_{d_{m-1}} \mid D^{m-2}_{d_{m-2}}, \ldots, D^1_{d_1}, I_2)}{\mathrm{P}(D^{m-1}_{d_{m-1}} \mid D^{m-2}_{d_{m-2}}, \ldots, D^1_{d_1}, I_2)} =$$

$$\frac{\sum_{r=0}^{m-1} \mathrm{P}(D^m_{d_m} \mid R^m_r, D^{m-1}_{d_{m-1}}, \ldots, D^{m-r}_{d_{m-r}}, I_2) \times \mathrm{P}(R^m_r, D^{m-1}_{d_{m-1}} \mid D^{m-2}_{d_{m-2}} \ldots, D^1_{d_1}, I_2)}{\mathrm{P}(D^{m-1}_{d_{m-1}} \mid D^{m-2}_{d_{m-2}} \ldots, D^1_{d_1}, I_2)}. \tag{29}$$

The reason for taking $D^{m-1}_{d_{m-1}}$ out of the conditional argument will be apparent in a moment. Note that the denominator is the probability of the previous observation given the previous data: in a recursive computation this would have been computed in the preceding step.

The products in the r sum contain the term given by (23), which is just the Johnson-Dirichlet conditional probability (18), and the joint probability for $(R^m_r, D^{m-1}_{d_{m-1}})$ given the previous data. This joint probability can be decomposed introducing R^{m-1}_s for the previous observation:

$$\mathrm{P}(R^m_r, D^{m-1}_{d_{m-1}} \mid D^{m-2}_{d_{m-2}}, \ldots, D^1_{d_1}, I_2) =$$

$$\frac{\sum_{s=1}^{m-1} \mathrm{P}(R^m_r, D^{m-1}_{d_{m-1}}, R^{m-1}_{s-1}, D^{m-2}_{d_{m-2}} \mid D^{m-3}_{d_{m-3}}, \ldots, D^1_{d_1}, I_2)}{\mathrm{P}(D^{m-2}_{d_{m-2}} \mid D^{m-3}_{d_{m-3}}, \ldots, D^1_{d_1}, I_2)} =$$

$$\frac{1}{\mathrm{P}(D^{m-2}_{d_{m-2}} \mid D^{m-3}_{d_{m-3}}, \ldots, D^1_{d_1}, I_2)} \sum_{s=1}^{m-1} \mathrm{P}(R^m_r \mid R^{m-1}_{s-1}, I_2) \times$$

$$\mathrm{P}(D^{m-1}_{d_{m-1}} \mid R^{m-1}_{s-1}, D^{m-2}_{d_{m-2}} \ldots, D^{m-s}_{d_{m-s}}, I_2) \times$$

$$\mathrm{P}(R^{m-1}_{s-1}, D^{m-2}_{d_{m-2}} \mid D^{m-3}_{d_{m-3}}, \ldots, D^1_{d_1}, I_2). \tag{30}$$

The last expression contains four kinds of terms. Let's analyse them.

- The denominator is the probability for the $(m-2)$th observation given the previous data. In a recursive algorithm this is known from two previous steps.
- The first factor within the s sum was simplified using the Markov property (26).
- The second factor within the s sum is given by the Johnson-Dirichlet model, according to discussion of eq. (23).
- The last factor within the s sum is analogous to the joint probability for $R_r^m, D_{d_{m-1}}^{m-1}$, which we are now calculating, but for the previous observation. In a recursive scheme it is known from the preceding step.

Combining formulae (29), (30), (32), (27), (18), and defining

$$A_m(d_m) =: \mathrm{P}(D_{d_m}^m \mid D_{d_{m-1}}^{m-1}, \dots, D_{d_1}^1, I_2), \tag{31a}$$

$$B_m(r, d_m) \coloneqq \mathrm{P}(D_{d_m}^m \mid R_r^m, D_{d_{m-1}}^{m-1}, \dots, D_{d_{m-r}}^{m-r}, I_2), \tag{31b}$$

$$C_m(r) \coloneqq \mathrm{P}(R_r^m, D_{d_{m-1}}^{m-1} \mid D_{d_{m-2}}^{m-2}, \dots, D_{d_1}^1, I_2), \tag{31c}$$

we find that the conditional probability we're seeking, $A_m(d_m)$, is given recursively by

$$A_m(d_m) = \frac{\sum_{r=0}^{m-1} B_m(r, d_m)\, C_m(r)}{A_{m-1}(d_{m-1})} \tag{32a}$$

$$B_m(r, d_m) = \frac{\Lambda\, \nu_{d_m} + r\, f_{d_m}}{\Lambda + r}, \qquad f_{d_m} \coloneqq \begin{Bmatrix}\text{rel. frequency of } d_m \text{ in}\\ \text{the past } r \text{ observations}\end{Bmatrix} \tag{32b}$$

$$C_m(r) = \frac{1}{A_{m-2}(d_{m-2})} \times \begin{cases} \sum_{s=1}^{m-1} h(s,m)\, B_{m-1}(s-1, d_{m-1})\, C_{m-1}(s-1) & \text{if } r = 0 \\ [1-h(r,m)]\, B_{m-1}(r-1, d_{m-1})\, C_{m-1}(r-1) & \text{if } r \geqslant 1 \end{cases} \tag{32c}$$

for all $d_m \in \{1, \dots, N\}$, $r \in \{0, \dots, m-1\}$. This recursive algorithm is summarized in table 2.

Some initial values are necessary for this recursive scheme. For $m = 1$, the first observation, the distribution is $\mathrm{P}(D_{d_1}^1 \mid I_2) = \nu_{d_1}$ by the Johnson-Dirichlet model, and trivially $\mathrm{P}(R_0^1 \mid I_2) = 1$. Owing to the certainty of R_0^1,

for $m = 2$ we have

$$\mathrm{P}(R_r^2, D_{d_1}^1 \mid I_2) = \mathrm{P}(R_r^2, R_0^1, D_{d_1}^1 \mid I_2) = \mathrm{P}(R_r^2 \mid R_0^1, I_2) \times \mathrm{P}(D_{d_1}^1 \mid I_2). \tag{33}$$

With the probabilities above from $m = 3$ onward the recursion can be calculated with eqs (32). These initialization steps are summarized in table 1.

4.3 Changepoint function and its uses

The changepoint function (27),

$$h(s, m) \coloneqq \mathrm{P}(R_0^m \mid R_s^{m-1}, I_2), \qquad m \in \{2, 3, \dots\}, \quad s \in \{1, \dots, m-1\} \tag{27}_{\mathrm{r}}$$

determines the robot's inference about the appearance of changepoints, or changes in the source of the data. It can represent a great variety of changepoint assumptions or beliefs.

For example, the belief that the data source is the more likely to change the longer it's been operative corresponds to a h that increases as s increases. The belief that the longer a data source operates the more likely it will persist corresponds to a h that decreases as s increases. Even more complex beliefs can be represented with the dependence on m. For example, we can have a shift from the first to the second belief above as m increases.

Our new robot I_2, thanks to this function, can reproduce a great variety of human inferential behaviours. How well can such robots mimic reproduce the inferential behaviour of the Plinko participants?

One way to assess this is to tune the stubbornness parameter Λ and the changepoint function h so that the robot's sequence of predictive distributions is as close to a specific participant's as possible. The reset-distribution parameter $\boldsymbol{\nu}$ can be set equal to the participant's initial distribution – thus assuming that the participant would again choose that distribution if he or she were informed about a change in the algorithm determining the outcomes.

This optimization leads to a possibly unique set of parameters $\big(\Lambda, h(s, m)\big)$ associated to a participant. We could then explore the distribution of these parameters across all participants, for example examining whether they cluster into few groups, and so on.

For such a study we need to choose a measure of discrepancy between two sequences of distributions, and to restrict our choice of changepoint

Table 1 Initial steps of predictive algorithm

1. Assign numerical values to Λ, ν
2. For $d_1 \in \{1, \dots, N\}$ calculate

$$A_1(d_1) \coloneqq \mathrm{P}(D^1_{d_1} \,|\, I_2) = \nu_{d_1}$$

3. Observe d_1
4. Keep $A_1(d_1)$ for the next two steps and $B_1(0) \equiv 1$ for the next step

5. For $r \in \{0, 1\}$:
   5.1. Calculate

$$C_2(r) \coloneqq \mathrm{P}(R^2_r, D^1_{d_1} \,|\, I_2) = A_1(d_1) \times \begin{cases} h(0,1) & \text{for } r = 0, \\ 1 - h(0,1) & \text{for } r = 1 \end{cases}$$

   5.2. For $d_2 \in \{1, \dots, N\}$, calculate

$$B_2(r, d_2) \coloneqq \mathrm{P}(D^2_{d_2} \,|\, R^2_r, D^1_{d_1}, I_2) = \frac{\Lambda\, \nu_{d_2} + r\, f_{d_2}}{\Lambda + r},$$

   where f_{d_2} is the relative frequency of outcome d_3 in the previous r observations
6. For $d_2 \in \{1, \dots, N\}$ calculate

$$A_2(d_2) \coloneqq \mathrm{P}(D^2_{d_2} \,|\, D^1_{d_1}, I_2) = \frac{\sum_{r=0}^{1} B_2(r, d_2) \times C_2(r)}{A_1(d_1)}$$

7. Observe d_2
8. Keep $A_2(d_2)$ for the next two steps, and $B_2(r, d_2)$, $C_2(r)$, $r \in \{0, 1\}$, for the next step
9. Set $m = 2$, go to step 1 of table 2

Table 2 Predictive algorithm

1. For $r \in \{0, \dots, m\}$:

   1.1. Calculate

$$C_{m+1}(r) \coloneqq \mathrm{P}(R_r^{m+1}, D_{d_m}^m \mid D_{d_{m-1}}^{m-1}, \dots, D_{d_1}^1, I_2) =$$

$$\frac{1}{A_{m-1}(d_{m-1})} \times$$

$$\begin{cases} \sum_{s=0}^{m-1} h(s,m) \times B_m(s, d_m) \times C_m(s) & \text{if } r = 0, \\ [1 - h(r-1, m)] \times B_m(r-1, d_m) \times C_m(r-1) & \text{if } r \geqslant 1 \end{cases}$$

   1.2. For $d_{m+1} \in \{1, \dots, N\}$, calculate

$$B_{m+1}(r, d_{m+1}) \coloneqq \mathrm{P}(D_{d_{m+1}}^{m+1} \mid R_r^{m+1}, D_{d_m}^m, \dots, D_{d_{m+1-r}}^{m+1-r}, I_2) = \frac{\Lambda \nu_{d_{m+1}} + r f_{d_{m+1}}}{\Lambda + r},$$

   where $f_{d_{m+1}}$ is the relative frequency of outcome d_{m+1} in the previous r observations

2. For $d_{m+1} \in \{1, \dots, N\}$, calculate

$$A_{m+1}(d_{m+1}) \coloneqq \mathrm{P}(D_{d_{m+1}}^{m+1} \mid D_{d_m}^m, \dots, D_{d_1}^1, I_2) = \frac{\sum_{r=0}^{m} B_{m+1}(r, d_{m+1}) \times C_{m+1}(r)}{A_m(d_m)}$$

3. Observe d_{m+1},
4. Keep $A_{m+1}(d_{m+1})$ for the next two steps, and $B_{m+1}(r, d_{m+1})$, $C_{m+1}(r)$, $r \in \{0, \dots, m\}$ for the next step
5. Increase m by 1, go to step 1

functions, to make the study computationally feasible. We now explore these two necessities.

4.4 Discrepancy between sequences of inferences

There are many measures of the discrepancy between two distributions: the relative entropy (also called Kullback-Leibler divergence or information discrimination) of the first distribution with respect to the second; or of the second with respect to the first; or the symmetrized version, namely the Shannon-Jansen discrepancy; or the Hellinger distance; and many others.

Given such a discrepancy for each of the M observations, we could define the total discrepancy as the average of these (L^1 norm), or the maximum among them (L^∞ norm), or other similar measures.

4.5 Restrictions on the changepoint function

With M observations, choosing a changepoint function corresponds to choosing $(M^2 - M)/2$ parameters with values in $[0, 1]$. We try to restrict this choice to make the optimization computationally feasible while maintaining an enough wide range of changepoint beliefs.

One restriction is to specify a logistic-polynomial changepoint function like this:

$$h(s, m; \gamma_0, \dots) = \sigma(\gamma_0 + \gamma_1 s + \gamma_2 m + \gamma_3 s^2 + \gamma_4 s m + \gamma_5 m^2 + \cdots) \quad (34)$$

where the logistic function

$$\sigma(x) \coloneqq \frac{1}{1 + \mathrm{e}^{-x}} \quad (35)$$

maps real numbers to $[0, 1]$.

In the rest of this study we choose a logistic-linear function:

$$h(s, m; \gamma, \gamma_\mathrm{s}, \gamma_\mathrm{m}) = \sigma\left(\gamma + \gamma_\mathrm{s} \frac{2s - M}{M} + \gamma_\mathrm{m} \frac{2m - M}{M}\right) \quad (36)$$

thus reducing our choice to a three-parameter family of changepoint functions. This family can reproduce, for example, the two simple kinds of changepoint beliefs mentioned in § 4.3, but not an m-dependent shift between them. Examples are shown in fig.***

✠ Luca: work in progress in the rest of this subsection

5 Notes on hierarchic models

5.1 How hierarchic models get updated

The probability calculus allows for inferences that learn from data in various degrees. We call 'model' a particular way of doing inference; each model is characterized by a particular capability of learning from data. These capabilities arise from a hierarchy of groups within groups of models.

At the bottom we have models that do not learn at all; we call them *independent*, because they assign independent probabilities to different data. For example, denoting a particular independent model by θ – which could be the value of a parameter identifying the model – and by I all other knowledge or assumptions besides this model, we have

$$\mathrm{p}(d_1, d_2 | \, \theta, I) = \mathrm{p}(d_1 | \, \theta, I)\, \mathrm{p}(d_2 | \, \theta, I) \tag{37}$$

for any two data d_1, d_2. This model does not learn because

$$\mathrm{p}(d_2 | \, d_1, \theta, I) = \mathrm{p}(d_2 | \, \theta, I), \tag{38}$$

that is, under this model one set of data is always irrelevant for the prediction of another set. The probability of an independent model given data d is

$$\mathrm{p}(\theta | \, d, I) = \frac{\mathrm{p}(d | \, \theta, I)\, \mathrm{p}(\theta | \, I)}{\sum_\theta \mathrm{p}(d | \, \theta, I)\, \mathrm{p}(\theta | \, I)}, \tag{39}$$

where $\mathrm{p}(\theta | \, I)$ is the probability over a range of such models based only on knowledge I, and $\mathrm{p}(d | \, \theta, I)$ is the *likelihood* of the independent model given the data.

We can introduce the capability of learning from data by considering a collection $\{\theta\}$ of independent models, each having a probability, and letting the data influence the probabilities of these models, rather than the model themselves.

This particular model, based on a collection of independent models, is usually called a *parametric* model. Let us denote a particular parametric model by μ. It is not independent because

$$\begin{aligned} \mathrm{p}(d_1, d_2 | \, \mu, I) &= \sum_\theta \mathrm{p}(d_1, d_2 | \, \theta, \mu, I)\, \mathrm{p}(\theta | \, \mu, I) \\ &\equiv \sum_\theta \mathrm{p}(d_1 | \, \theta, \mu, I)\, \mathrm{p}(d_2 | \, \theta, \mu, I)\, \mathrm{p}(\theta | \, \mu, I), \end{aligned} \tag{40}$$

which doesn't factorize unless $\mathrm{p}(\theta|\,\mu, I)$ is a delta. The first equality comes from the law of total probability. Such a model learns because

$$\mathrm{p}(d_2|\,d_1, \mu, I) = \sum_\theta \mathrm{p}(d_2|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,d_1, \mu, I) \tag{41a}$$

with

$$\mathrm{p}(\theta|\,d_1, \mu, I) = \frac{\mathrm{p}(d_1|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,\mu, I)}{\sum_\theta \mathrm{p}(d_1|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,\mu, I)}, \tag{41b}$$

where we see that data d_1 affect not the probability of d_2 directly, but the probability distribution for the various independent models. The probability of a parametric model given data d is

$$\mathrm{p}(\mu|\,d, I) = \frac{\mathrm{p}(d|\,\theta, I)\,\mathrm{p}(\mu|\,I)}{\sum_\mu \mathrm{p}(d|\,\mu, I)\,\mathrm{p}(\mu|\,I)} \tag{42a}$$

with

$$\mathrm{p}(d|\,\mu, I) = \sum_\theta \mathrm{p}(d|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,\mu, I). \tag{42b}$$

where $\mathrm{p}(\mu|\,I)$ is the probability over a range of parametric models based only on knowledge I. The last expression (42b) is the *likelihood* of the model given the data, and we see that it's given by a mixture of independent models.

Equations (42) and (40) show that a parametric model is constructed as an uncertainty over independent models, and eq. (41) shows that data affect this latter uncertainty. It is as if we were considering different ways of doing inference, and inferring which of such inferences is most probable. Each bottom inference is incapable to learn from data, but our inferences about these inferences can learn from data.

We can proceed analogously and consider a collection $\{\mu\}$ of parametric models, each having a probability, and letting the data influence this probability as well. The model constructed this way is usually called a one-level hierarchic model – even though we've seen that a parametric model can also be considered as hierarchic. Let us denote such a model by χ. The probability of the data is

$$\begin{aligned}\mathrm{p}(d_1, d_2|\,\chi, I) &= \sum_\mu \mathrm{p}(d_1, d_2|\,\mu, \chi, I)\,\mathrm{p}(\mu|\,\chi, I)\\ &\equiv \sum_\mu \Big\{\sum_\theta \Big[\prod_i \mathrm{p}(d_i|\,\theta, \mu, \chi, I)\Big]\,\mathrm{p}(\theta|\,\mu, \chi, I)\Big\} \times\\ &\qquad\qquad \mathrm{p}(\mu|\,\chi, I).\end{aligned} \tag{43}$$

Learning takes place this way:

$$\mathrm{p}(d_2|\, d_1, \chi, I) = \sum_{\mu} \Big[\sum_{\theta} \mathrm{p}(d_2|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, d_1, \mu, \chi, I) \Big]\, \mathrm{p}(\mu|\, d_1, \chi, I) \tag{44a}$$

with

$$\begin{aligned} \mathrm{p}(\theta|\, d_1, \mu, \chi, I) &= \frac{\mathrm{p}(d_1|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, \mu, \chi, I)}{\sum_{\theta} \mathrm{p}(d_1|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, \mu, \chi, I)} \\ &= \frac{\mathrm{p}(d_1|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, \mu, \chi, I)}{\mathrm{p}(d_1|\, \mu, \chi, I)}, \end{aligned} \tag{44b}$$

$$\mathrm{p}(\mu|\, d_1, \chi, I) = \frac{\mathrm{p}(d_1|\, \mu, \chi, I)\, \mathrm{p}(\mu|\, \chi, I)}{\sum_{\mu} \mathrm{p}(d_1|\, \mu, \chi, I)\, \mathrm{p}(\mu|\, \chi, I)}. \tag{44c}$$

Note that the denominator of the update formula (44b) for θ is the updated probability (44c) for μ. Note also that the space of independent models $\{\theta\}$ can be different for different μ.

With a hierarchic model, it is as if we were considering different 'super-inferences' about ways of doing inferences, and inferring which of such super-inferences is the most probable. This model learns from the data in two ways: the data first give more probability to one or another parametric model, and then give more probability to one or another independent model within that parametric model. From another point of view we can say that the data perform first a coarsen selection, and then a finer one within each coarser selection.

We can of course multiply this kind of hierarchy ad libitum, proceeding as we've done so far.

5.2 Flattening hierarchic models

The subdivision of learning into two or more levels of different coarseness can be very convenient, but mathematically it's always equivalent to one single subdivision at the finest level. In other words, any hierarchic model can always be rewritten as a parametric one. Let's see how, in the case of a hierarchic model like (43).

We said that each parametric model μ has a set $\{\theta\} = \{\theta\}_\mu$ of underlying independent models. For example, in the case of real-valued data, one set could contain normal distributions with the same variance and different means; another set could contain uniform distributions with different supports; yet another set could contain Cauchy distributions with the same location parameter and different scale parameters. These sets can be pairwise disjoint, overlapping, or even identical for different μ.

First of all let's consider each such set $\{\theta\}_\mu$ as formally distinct from all others for different μ. We consider the union of all these sets, denoting a member of this union by Θ:

$$\{\Theta\} \coloneqq \bigcup_\mu \{\theta\}_\mu. \tag{45}$$

Now consider the predictive probability (43) for the hierarchic model:

$$\mathrm{p}(d\,|\,\chi, I) = \sum_\mu \Big[\sum_\theta \mathrm{p}(d\,|\,\theta, \mu, \chi, I)\,\mathrm{p}(\theta\,|\,\mu, \chi, I)\Big]\,\mathrm{p}(\mu\,|\,\chi, I). \tag{46}$$

If we decree that $\mathrm{p}(\Theta\,|\,\mu, \chi, I) = 0$ if $\Theta \notin \{\theta\}_\mu$, we can extend the sum over θ (for fixed μ) over all Θ. Moreover, since Θ contains information about μ, the latter becomes irrelevant in the conditional of the probability $\mathrm{p}(d\,|\,\theta, \mu, \chi, I)$. The predictive probability above then becomes

$$\begin{aligned} \mathrm{p}(d\,|\,\chi, I) &= \sum_\Theta \mathrm{p}(d\,|\,\Theta, \chi, I) \sum_\mu \mathrm{p}(\Theta\,|\,\mu, \chi, I)\,\mathrm{p}(\mu\,|\,\chi, I), \\ &= \sum_\Theta \mathrm{p}(d\,|\,\Theta, \chi, I)\,\mathrm{p}(\Theta\,|\,\chi, I), \end{aligned} \tag{47}$$

where the last equality follows from the law of total probability. What's important in the last formula is that the probability $\mathrm{p}(d\,|\,\Theta, \chi, I)$ factorizes over conjunctions of data; that is, it is an independent model. The model above is therefore just a parametric model.

The last step consists in joining together into a single value all those values of Θ which lead to identical predictive distributions $\mathrm{p}(d\,|\,\Theta, \chi, I)$. The probability $\mathrm{p}(\Theta\,|\,\chi, I)$ for such a value will be the sum of the probability for the various equivalent values.

✠ to be cont'd

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Adams, R. P., MacKay, D. J. C. (2007): *Bayesian online changepoint detection*. http://hips.seas.harvard.edu/content/bayesian-online-changepoint-detection.

Bayes, T. (1763): *An essay towards solving a problem in the doctrine of chances*. Phil. Trans. R. Soc. Lond. **53**, 370–418. http://www.stat.ucla.edu/history/essay.pdf; with an introduction by Richard Price. The *Scholium* of p. 392 is reprinted and analysed in Stigler (1982).

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, reprint. (Wiley, New York). First publ. 1994.

Chatfield, C. (1995): *Model uncertainty, data mining and statistical inference*. J. Roy. Stat. Soc. A **158**[3], 419–444. See also discussion in Copas, Davies, Hand, Lunneborg, Ehrenberg, Gilmour, Draper, Green, et al. (1995).

Copas, J. B., Davies, N., Hand, D. J., Lunneborg, C. E., Ehrenberg, A. S., Gilmour, S. G., Draper, D., Green, P. J., et al. (1995): *Discussion of the paper by Chatfield [Model uncertainty, data mining and statistical inference]*. J. Roy. Stat. Soc. A **158**[3], 444–466. See Chatfield (1995).

Daft Punk (2005a): *Human after all*. In: Daft Punk (2005b).

— (2005b): *Human After All*. (Virgin, worldwide).

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford).

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013), ch. 2, 19–29.

de Finetti, B. (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. **IV**[5], 86–133. http://www.brunodefinetti.it/Opere.htm.

— (1937): *La prévision : ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**[1], 1–68. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

DeGroot, M. H. (2004): *optimal statistical decisions*, reprint. (Wiley, New York).

Diaconis, P., Ylvisaker, D. (1979): *Conjugate priors for exponential families*. Ann. Stat. **7**[2], 269–281.

Filipowicz, A., Valadao, D., Anderson, B., Danckert, J. (2014): *Measuring the influence of prior beliefs on probabilistic estimations*. Proc. Annu. Meet. Cogn. Sci. Soc. **36**, 2198–2203.

— (2016): *Rejecting outliers: surprising changes do not always improve belief updating*. Decision ******[**]**, **.

Good, I. J. (1965): *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. (MIT Press, Cambridge, USA).

Hailperin, T. (1996): *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. (Associated University Presses, London).

— (2011): *Logic with a Probability Semantics: Including Solutions to Some Philosophical Problems*. (Lehigh University Press, Plymouth, UK).

Heath, D., Sudderth, W. (1976): *De Finetti's theorem on exchangeable variables*. American Statistician **30**[4], 188–189.

iso (2006): *ISO 3534-1:2006: Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization.

— (2009): *ISO 80000:2009: Quantities and units*. International Organization for Standardization. First publ. 1993.

Jaynes, E. T. (1996): *Monkeys, kangaroos, and N*. http://bayes.wustl.edu/etj/node1.html. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).

— (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html, http://omega.albany.edu:8008/JaynesBook.html.

Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). https://archive.org/details/logic03john.

— (1932): *Probability: the deductive and inductive problems*. Mind **41**[164], 409–423. With some notes and an appendix by R. B. Braithwaite.

Keynes, J. M. (2013): *A Tract on Monetary Reform*, repr. of second ed. (Cambridge University Press, Cambridge). First publ. 1923.

Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.

Laplace (Marquis de), P. S. (1819): *Essai philosophique sur les probabilités*, 4th ed. (Courcier, Paris). Repr. as the Introduction of Laplace (1820), pp. V–CLIII; http://gallica.bnf.fr/document?O=N077595.

— (1820): *Théorie analytique des probabilités*, 3rd ed. (Courcier, Paris). First publ. 1812; repr. in Laplace (1886); http://gallica.bnf.fr/document?O=N077595.

— (1886): *Œuvres complètes de Laplace. Tome septième : Théorie analytique des probabilités*. (Gauthier-Villars, Paris). 'Publiées sous les auspices de l'Académie des sciences, par MM. les secrétaires perpétuels'; http://gallica.bnf.fr/notice?N=FRBNF30739022.

Nassar, M. R., Wilson, R. C., Heasly, B., Gold, J. I. (2010): *An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment*. J. Neurosci. **30**[37], 12366–12378. https://www.princeton.edu/~rcw2/publications.html.

Stigler, S. M. (1982): *Thomas Bayes's Bayesian inference*. J. Roy. Stat. Soc. A **145**[2], 250–258.

Zabell, S. L. (1982): *W. E. Johnson's "sufficientness" postulate*. Ann. Stat. **10**[4], 1090–1099. Repr. in Zabell (2005 pp. 84–95).

— (2005): *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. (Cambridge University Press, Cambridge).