Bayesian Plinko

Study notes

Alex, Luca, Yasser, Britt, James

14 April 2018; updated 1 July 2018

<div align="right">

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

</div>

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 behaviour is the 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 be 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

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_d^i. In this context, choosing a statistical model I is equivalent to choosing the numerical values of the joint distribution

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

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, $P(D_{d_1}^1 \mid I)$, the distributions for the second observation conditional on the first, $P(D_{d_2}^2 \mid D_{d_1}^1, I)$, and so on up to the distributions for the Mth observation conditional on the rest, $P(D_{d_M}^M \mid D_{d_{M-1}}^{M-1}, \ldots, D_{d_1}^1, I)$. 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 condizionalization and marginalization.

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] 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

---

[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).

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 described above, the set of possible robots is a simplex of $N^M - 1$ dimensions, since this is the set of all possible distributions for M joint outcomes, each having N possible values. In the case of the Plinko experiments (Filipowicz et al. 2014; 2016) it has $40^{200} - 1 \approx 3 \times 10^{320}$ dimensions.

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.

3   First study: exchangeable-model robot

3.1   The first Bayesian robot

We first consider a robot programmed with either of these two equivalent assumptions, denoted I:

- The joint probability distribution (1) for *any* number of observations is invariant with respect to exchanges of their order. That is, the order of the observations is irrelevant for making inferences.

- Only the relative frequencies of past observations are relevant for making inferences. Any additional data about known observations is irrelevant and can be discarded.

In either of these assumptions it's implicit that probability distributions for different number of observations must of course be consistent with one another through marginalization.

These two equivalent assumptions are 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 (references above) shows that the assumption of infinite exchangeability implies that the probability distribution (1) must have this mathematical form, for any m:

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

where q is a normalized N-tuple of positive numbers:

$$\Delta := \{q \in \mathbf{R}^N \mid q_i \geqslant 0, \textstyle\sum_{i=1}^{N} q_i = 1\}, \tag{3}$$

and $p(q \mid I)\, dq$ is a normalized density function, called *prior parameter density*.

The N-tuple q can be interpreted as the relative long-run frequencies of the possible outcomes[2], and $p(q \mid I)\, dq$ as their probability density. 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, not knowing their particular order in the observation, assigns to the occurrence of each a probability proportional to its frequency: this is the term $\prod_{i=1}^{m} q_{d_i}$ in the integral. Then, being uncertain about the long-run frequencies, the robot assigns to them the density $p(q \mid I)\, dq$. The combined uncertainty about the observations is then eq. (2) 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 $p(q \mid I)\, dq$. Once this is chosen – for example, the constant density $(N - 1)!\, dq$, 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 this part of the assumptions I.

In fact, the robot resulting from the assumptions I 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 (2) with $N = 40$ is

$$P(D_{37}^1, D_6^2, D_{25}^3, D_{37}^4 D_{19}^5 \mid I) = \int_\Delta q_6\, q_{19}\, q_{25}\, q_{37}^{\,2}\, p(q \mid I)\, dq. \tag{4}$$

In the following we omit the integration domain Δ.

3.2   Probability updates from observations

Using Bayes's theorem the robot can calculate from (2) the probability distribution for the mth observation, having observed outcomes $d_1, \ldots, d_{m-1}$,

$$P(D_{d_m}^m \mid D_{d_{m-1}}^{m-1}, \ldots, D_{d_2}^2, D_{d_1}^1, I), \tag{5}$$

---

[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)

obtaining

$$P(D_{d_m}^m | D_{d_{m-1}}^{m-1}, \ldots, D_{d_2}^2, D_{d_1}^1, I) = \int q_{d_m} p(q | D_{d_1}^1, \ldots, D_{d_{m-1}}^{m-1}, I) \, dq, \quad (6a)$$

with

$$p(q | D_{d_1}^1, \ldots, D_{d_{m-1}}^{m-1}, I) = \frac{\left(\prod_{i=1}^{m-1} q_{d_i}\right) p(q | I)}{\int \left(\prod_{i=1}^{m-1} q'_{d_i}\right) p(q' | I) \, dq'}. \quad (6b)$$

For our numeric example (4) this is

$$P(D_{19}^5 | D_{37}^1, D_6^2, D_{25}^3, D_{37}^4, I) = \int q_{19} \, p(q | D_{37}^1, D_6^2, D_{25}^3, D_{37}^4, I) \, dq, \quad (7a)$$

$$p(q | D_{37}^1, D_6^2, D_{25}^3, D_{37}^4, I) = \frac{q_6 \, q_{25} \, q_{37}^2 \, p(q | I)}{\int q'_6 \, q'_{25} \, q'^2_{37} \, p(q' | I) \, dq'}. \quad (7b)$$

Once the density $p(q | I) \, dq$ is chosen, the robot can perform the Plinko task just like all others participants.

3.3 General remarks on the robot's behaviour

The infinite-exchangeability formula (6) leads 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 $p(q | D, I) \, dq$ will become 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 $p(q | I) \, dq$ – 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 q' will require a lot of data to move the predictions to a very different frequency q''.

- Suppose that robot receives a long sequence of observations concentrated around frequencies q' – say, a very long sequence of 3s in a row – and then another long sequence concentrated around different frequencies 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 data tell us each participant's probabilities for the first observation,

$$\mathrm{p}(D_d^1 | \text{participant}) \equiv \int q_d \, \mathrm{p}(q | \text{participant}) \, \mathrm{d}q, \qquad (8)$$

but not their prior parameter density $\mathrm{p}(q | \text{participant}) \, \mathrm{d}q$.

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

$$\mathrm{p}(q | I) = \frac{\Gamma(\Lambda)}{\prod_i \Gamma(\Lambda \, \nu_i)} \prod_{i=1}^{N} q_i^{\Lambda \nu_i - 1}, \qquad \Lambda > 0, \nu \in \Delta. \qquad (9)$$

The N independent parameters (Λ, ν) 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 $f := (f_1, \dots, f_N)$ are the observed outcomes' relative frequencies, this assumption can be mathematically expressed as

$$\mathrm{P}(D_d^{m+1} | f, N, I) = \mathrm{P}(D_d^{m+1} | f_d, N, I) \qquad \text{for all } d \in \{1, \dots, N\}. \quad (10)$$

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.

The density (9) 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

$$\int_\Delta \prod_{i=1}^N q_i^{x_i-1} \, \mathrm{d}q = \frac{\prod_i \Gamma(x_i)}{\Gamma(\sum_i x_i)}. \tag{11}$$

In fact, inserting this prior in the infinite-exchangeability formula (6) we find the joint distribution (1), for any m:

$$p(D_{d_1}^1, \dots, D_{d_m}^m \mid I) = \frac{\Gamma(\Lambda)}{\Gamma(\Lambda + m)} \prod_{d=1}^N \frac{\Gamma(\Lambda \nu_d + m f_d)}{\Gamma(\Lambda \nu_d)}, \tag{12}$$

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

In particular, the initial distribution is simply

$$P(D_d^1 \mid I) = \nu_d. \tag{13}$$

This formula says that the Johnson-Dirichlet prior can produce any possible initial probability distribution, by appropriately choosing $\boldsymbol{\nu}$. Thus the robot can be programmed to make the same initial inference as a participant, just by setting its parameters $\boldsymbol{\nu}$ equal to the participant's initial distribution. The parameter Λ cannot be directly determined by a participant's inferences. We call this parameter the *stubbornness* of the robot; here's the reason.

From eqs (6b), (9), and (11) we can calculate the parameter density updated after the observations $1, \dots, m-1$:

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

$$\text{with} \quad \Lambda' = \Lambda + m - 1, \quad \boldsymbol{\nu}' = \frac{\Lambda \boldsymbol{\nu} + (m-1)\boldsymbol{f}}{\Lambda + m - 1} \tag{14}$$

where $\boldsymbol{f}$ are the relative frequencies of the outcomes in the $m-1$ observations. This formula, together with (6) and (11), yields the robot's inference for the mth observation conditional on the previous ones:

$$P(D_d^m \mid D_{d_{m-1}}^{m-1}, \dots, D_{d_1}^1, I) = \frac{\Lambda \nu_d + (m-1) f_d}{\Lambda + m - 1}. \tag{15}$$

The latter formula has an interesting interpretation: the distribution the robot calculates for the mth observation is equal to a weighted sum

of the initial distribution (13) and the observed relative frequencies f. The weights are the parameter Λ and the number of observations.

Suppose that after some observations the predictive distribution is $\boldsymbol{\nu}$, and that the next outcome is k. Then the probability distribution for the slots is updated to:

$$P(D_j^{m+1} | D_k^m, \boldsymbol{\nu}^m, \Lambda^m, I) = \nu_j^{m+1} := \frac{\Lambda^m}{\Lambda^m + 1} \nu_j^m + \begin{cases} \dfrac{1}{\Lambda^m + 1} & \text{if } j = k, \\ 0 & \text{if } j \neq k, \end{cases}$$

$$\text{and } \Lambda^{m+1} := \Lambda^m + 1. \quad (16)$$

This update corresponds to a participant's raising the bar assignment under slot k, leaving the others untouched, and/or lowering the bar assignments for *all* other slots by the same proportion. The parameter Λ is increased by 1. The larger Λ, the more reluctant the robot is in revising its guesses in the light of new observations. The update formula (14) says that the robot behaves as if it had already made Λ observations with outcome frequencies $\boldsymbol{\nu}$.

3.5 Examples: participant vs robot

Let's choose a participant, and use formula (13) to choose the $\boldsymbol{\nu}$ parameters of the robot's prior, equating it to the initial predictive distribution of the participant. Let's set a value for the robot's stubbornness Λ, and check how the robot updates its predictive distribution, using formula (14), while observing the same outcomes as the participant.

Figure 1 shows the means and standard deviations of the sequence such predictive 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 variability in the figure shows. The program generating the outcomes had a change in standard deviation, shifting to a narrower distribution at trial 101. The robot adapted to this change very slowly.

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.
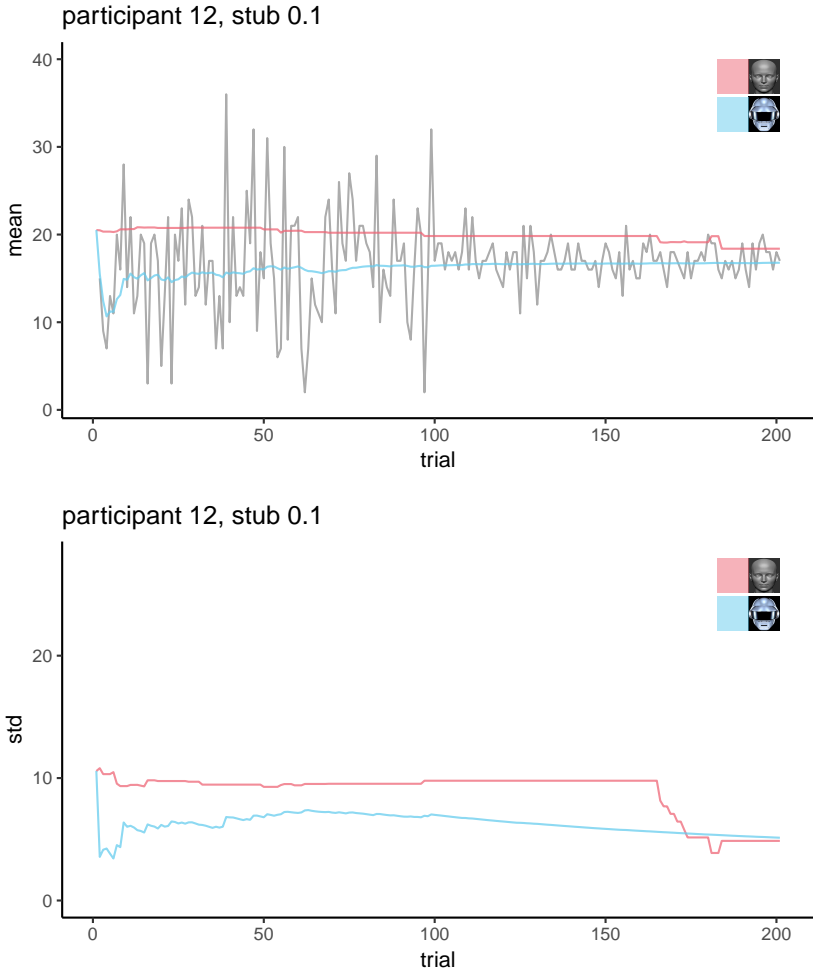
Figure 1   Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubbornness $\Lambda = 0.1$
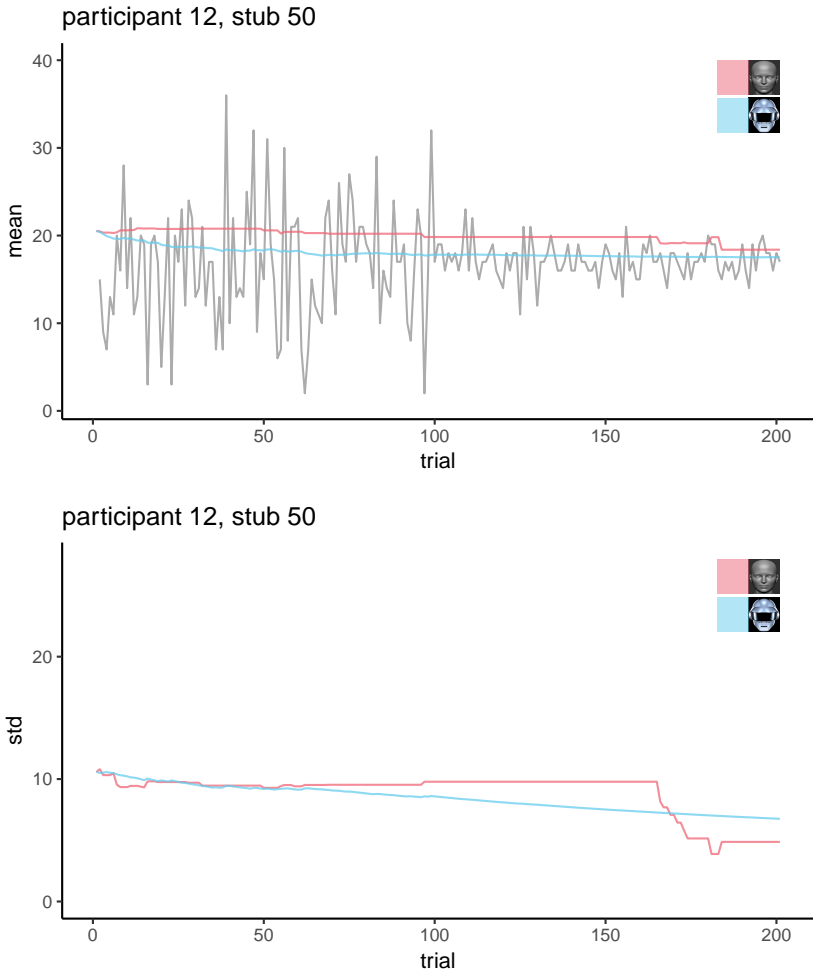
Figure 2   Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubbornness $\Lambda = 50$
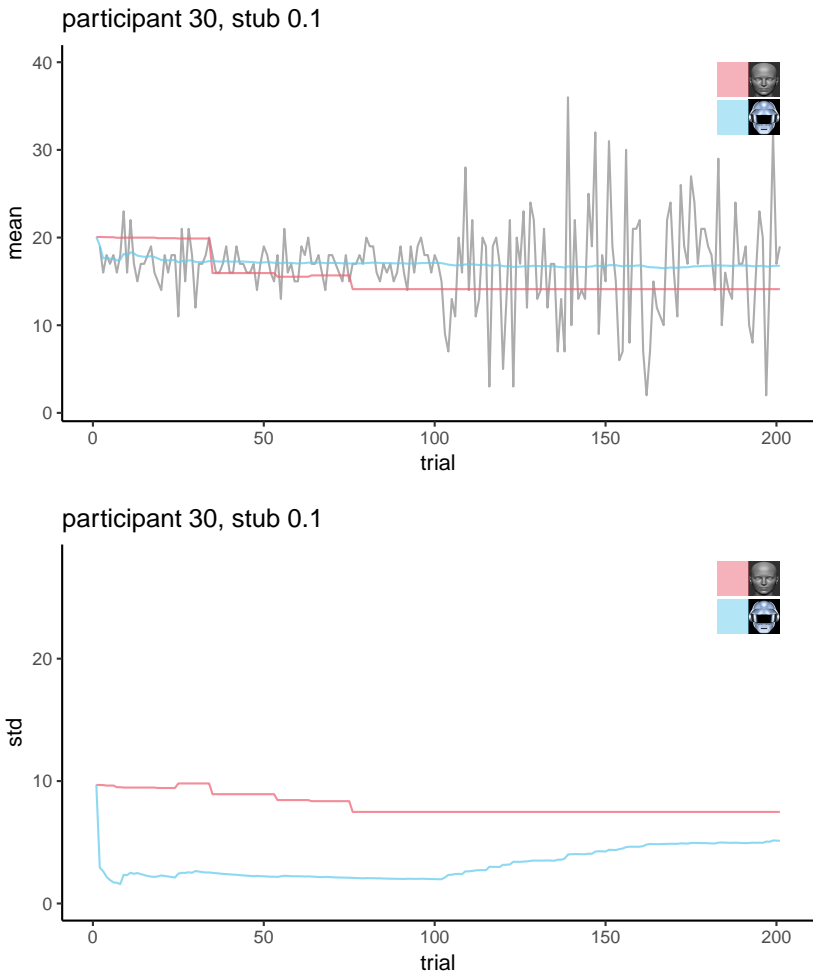
Figure 3 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubbornness $\Lambda = 0.1$
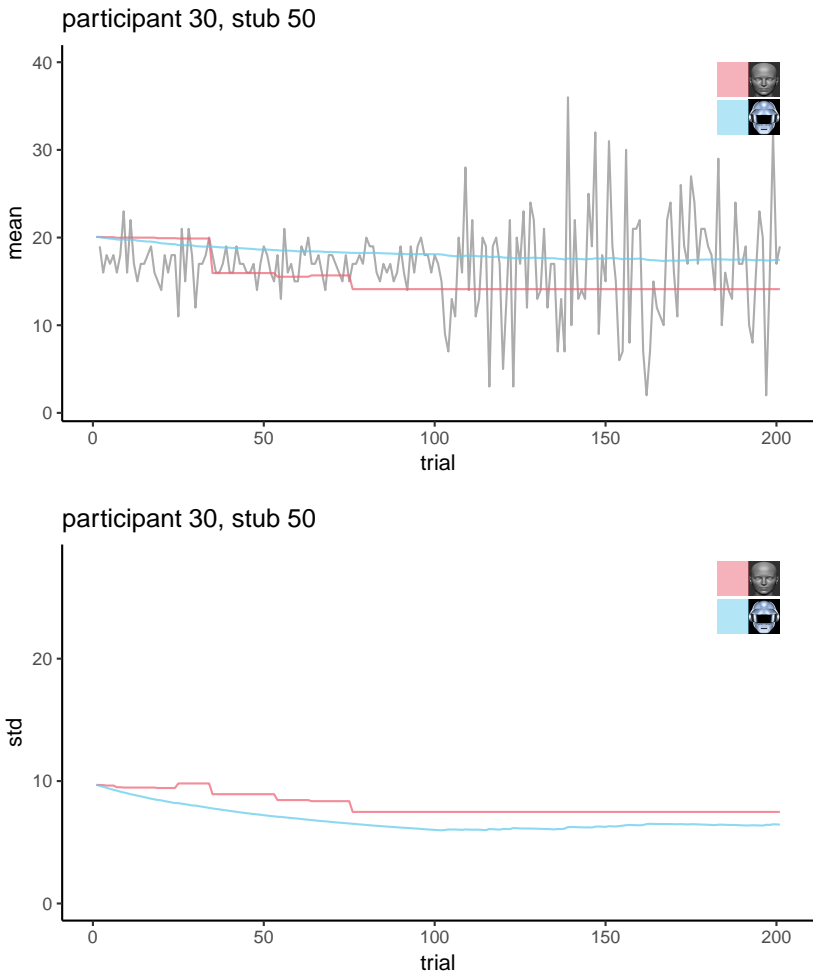
Figure 4 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubbornness $\Lambda = 50$

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 trial 100. If we could plot these outcomes while they are generated, we would likely notice the change by around trial 25. Our robot, however, can't detect this change for the reasons explained in § 3.3; any outcomes from narrow or wide generating processes are mingled in the robot's memory.

Only non-exchangeable or hierarchic models can exhibit a short evidence memory and be 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.

• A participant who, after observing outcome k, raises the bar under that slot *and nearby bars* is therefore not acting according to a Johnson-Dirichlet exchangeable model.

• Infinitely exchangeable priors are incapable of quickly adapting to changes in the empirical statistics of the outcomes.

3.6 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, at trial 101 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 trial 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; 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. ⚑ L: so good?! must recheck the script.
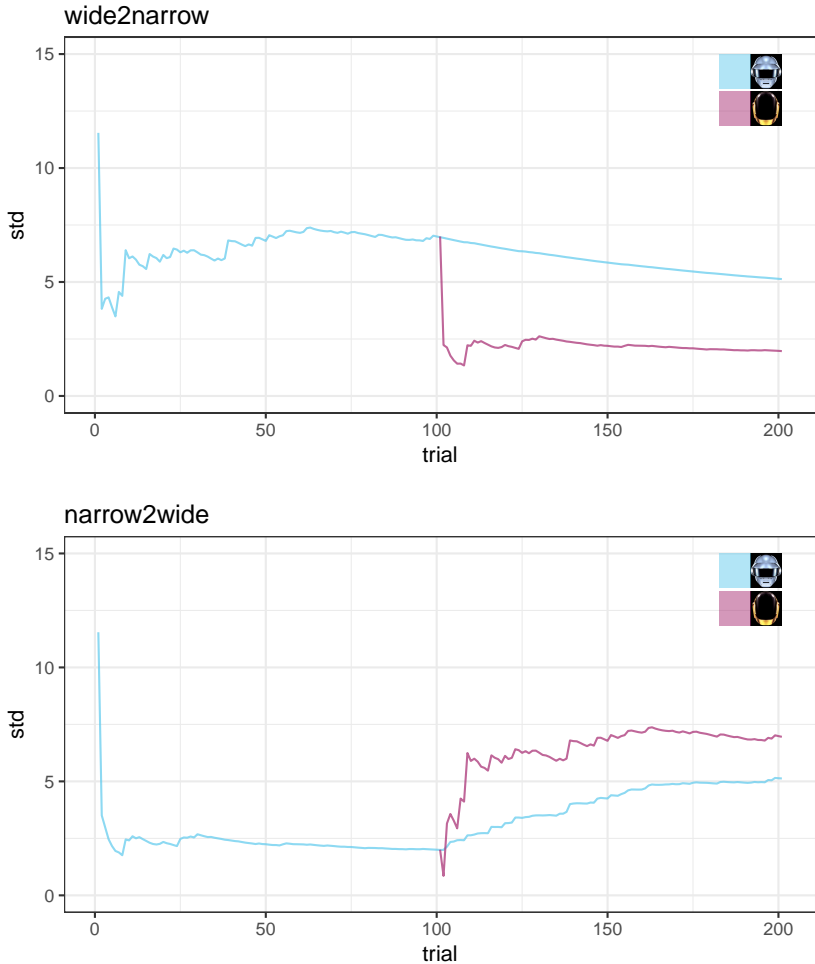
Figure 5   Adaptation speed of the standard deviation for a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change. The final relative entropy for the first robot relative to the second is 1.01 in the wide-to-narrow case, vs 0.260 in 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
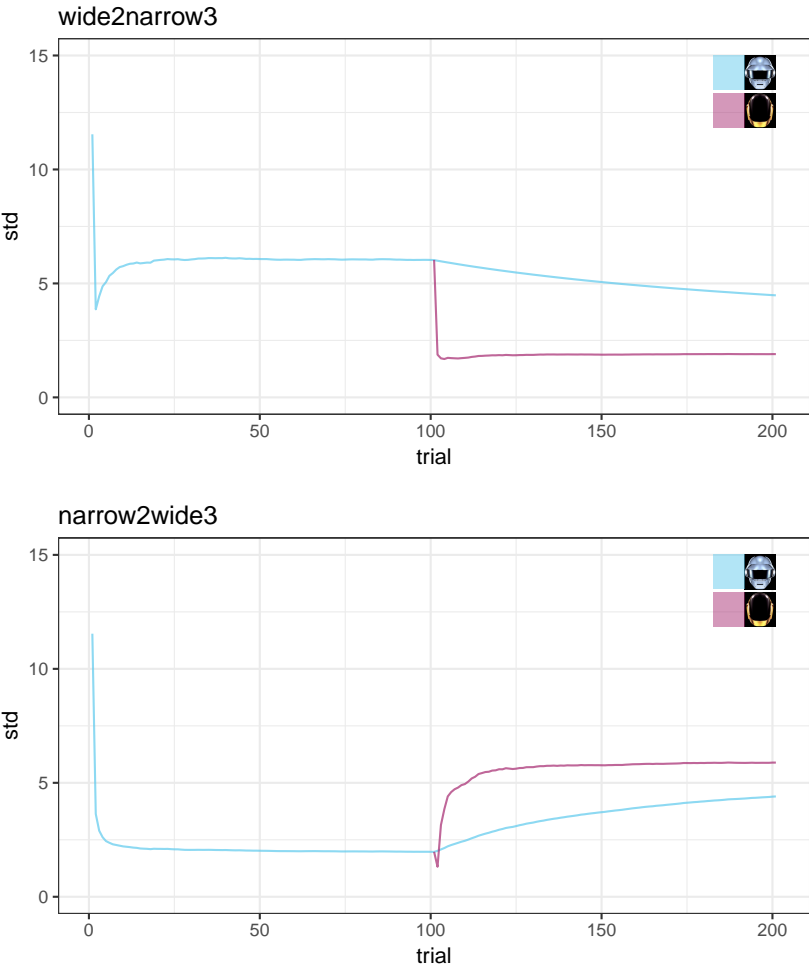
Figure 6   Adaptation speed of the standard deviation for a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change, averaged over 100 experiments. The normals have same mean 17 and standard deviations 6 and 1.9.
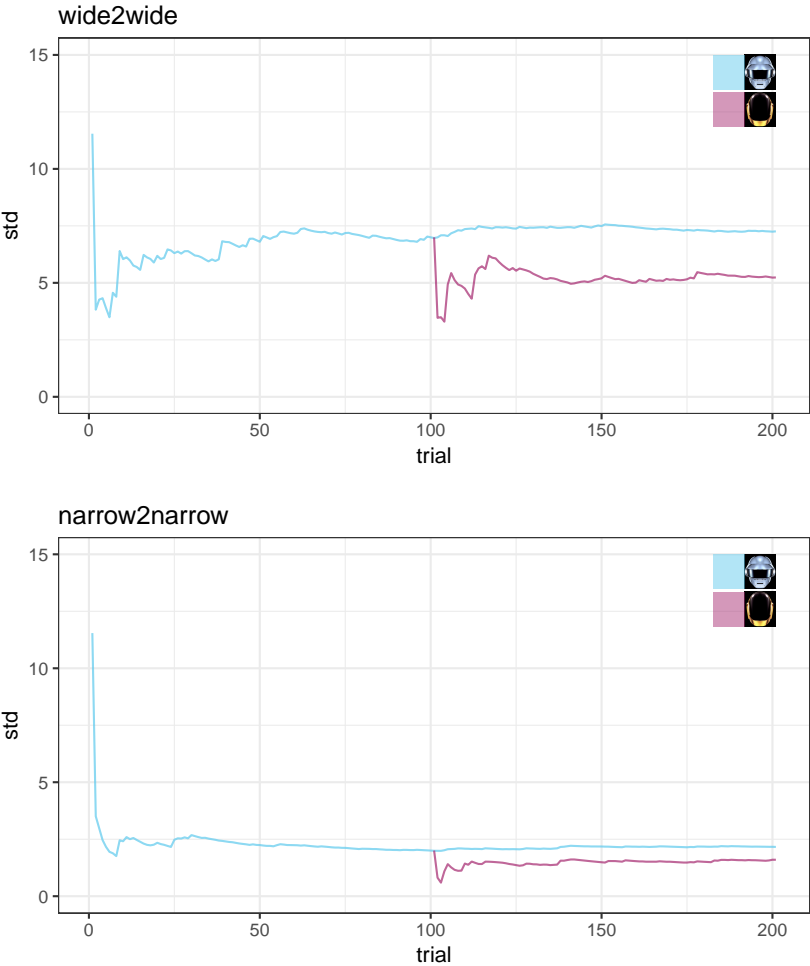
Figure 7 Adaptation speed of a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change.

4 Second study: exchangeable model with change-points

4.1 The robot looks for change-points

The robot of the previous study remembers past observations 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 order of observations into account in a simple way.

This new robot assumes that the sequence of observations is divided into subsequences of contiguous observations. *Within each subsequence* the order of the observations doesn't matter: frequencies are sufficient statistics, and the robot therefore uses an infinitely exchangeable model with a Johnson-Dirichlet parameter density, as in the previous study. Only observations within the subsequence contribute to the update of the parameter density. Observations from other subsequences are irrelevant to inferences within each subsequence, provided that the start point of that subsequence is known. The stubbornness and frequency parameters are therefore 'reset' at the start of each subsequence.

The observations at which the subsequences start, called change-points, are guessed from the whole sequence of observations. The robot is effectively considering all possible numbers and positions of change-points. For example, every observation could be a subsequence by itself, or all observations could belong to just one subsequence – which was the assumption of the robot of the previous study. According to Bayes's theorem, the robot assigns a probability to each possible division pro-portionally to a prior times the product of the probabilities for the data in each subsequence determined by that division, calculated with the Johnson-Dirichlet model described above:

P(particular division into subsequences| data, I) $\propto$

$$\text{\textit{Johnson-Dirichlet}} \begin{cases} \text{P(data in 1st subsequence| particular division, } I) \\ \times \text{P(data in 2nd subsequence| particular division, } I) \\ \times \cdots \end{cases}$$

$$\times \text{P(particular division| } I), \quad (17)$$

where I stands for other background assumptions. The final inference is a combination of the inferences conditional on each possible division:

P(new observation| data, I) $\propto$

$$\sum_{\text{all divisions}} \text{P(new observation| division, data, } I) \times \text{P(division| data, } I), \quad (18)$$

where the probability for the new observation given the division is determined by the Johnson-Dirichlet model.

Adams & MacKay (2007) give an algorithm to do the calculations above on-line as new observations are made. Before each observation the algorithm saves the probability distribution for the divisions (17), which is used to calculate the updated distribution with the new observation adjoined to the data.

For each observation i, consider the proposition 'only the previous r observations are relevant for the ith one', denoted by R_r^i. This proposition is equivalent to say that observations $\{i - r, i - r + 1, \ldots, i\}$ form a subsequence. Obviously $0 \leqslant r \leqslant i - 1$, where $i - 1$ is the total number of observations before the ith one; if $r = 0$ then the ith observation is a changepoint, starting a new subsequence, and no past data are relevant for its inference. In symbols

$$\text{P}(D_{d_{m+1}}^{m+1}| R_r^{m+1}, D_{d_m}^m, \ldots, D_{d_1}^1, I) =$$

$$\begin{cases} \text{P}(D_{d_{m+1}}^{m+1}| R_r^{m+1}, D_{d_m}^m, \ldots, D_{d_{m+1-r}}^{m+1-r}, I) & \text{if } 1 \leqslant r \leqslant m, \\ \\ \text{P}(D_{d_{m+1}}^{m+1}| R_r^{m+1}, I) & \text{if } r = 0, \end{cases}$$

with probabilities given by the Johnson-Dirichlet model (6), (9), (16).
$$(19)$$

The probability for the $(m + 1)$th observation given the data can be thus decomposed

$$P(D_{d_{m+1}}^{m+1}| D_{d_m}^m, \ldots, D_{d_1}^1, I) = \frac{P(D_{d_{m+1}}^{m+1}, D_{d_m}^m | D_{d_{m-1}}^{m-1}, \ldots, D_{d_1}^1, I)}{P(D_{d_m}^m | D_{d_{m-1}}^{m-1}, \ldots, D_{d_1}^1, I)} =$$

$$\frac{\sum_{r=0}^{m} P(D_{d_{m+1}}^{m+1}| R_r^{m+1}, D_{d_m}^m, \ldots, D_{d_{m+1-r}}^{m+1-r}, I) \times P(R_r^{m+1}, D_{d_m}^m | D_{d_{m-1}}^{m-1} \ldots, D_{d_1}^1, I)}{P(D_{d_m}^m | D_{d_{m-1}}^{m-1} \ldots, D_{d_1}^1, I)}.$$

$$(20)$$

The reason for this taking $D_{d_m}^m$ out of the conditional 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 sum contain the term given by (19) and the joint probability for $R_r^{m+1}, D_{d_m}^m$ given the previous data. This joint probability can be decomposed introducing R_s^m for the previous observation:

$$P(R_r^{m+1}, D_{d_m}^m | D_{d_{m-1}}^{m-1}, \ldots, D_{d_1}^1, I) =$$

$$\frac{\sum_{s=0}^{m-1} P(R_r^{m+1}, D_{d_m}^m, R_s^m, D_{d_{m-1}}^{m-1}| D_{d_{m-2}}^{m-2}, \ldots, D_{d_1}^1, I)}{P(D_{d_{m-1}}^{m-1}| D_{d_{m-2}}^{m-2}, \ldots, D_{d_1}^1, I)} =$$

$$\frac{1}{P(D_{d_{m-1}}^{m-1}| D_{d_{m-2}}^{m-2}, \ldots, D_{d_1}^1, I)} \sum_{s=0}^{m-1} P(R_r^{m+1}| R_s^m, I) \times$$

$$P(D_{d_m}^m | R_s^m, D_{d_{m-1}}^{m-1} \ldots, D_{d_{m-s}}^{m-s}, I) \times$$

$$(21)$$

$$P(R_s^m, D_{d_{m-1}}^{m-1}| D_{d_{m-2}}^{m-2}, \ldots, D_{d_1}^1, I).$$

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

1. The denominator is the probability for the $(m - 1)$th observation given the previous data. In a recursive algorithm this has been calculated in a previous step.

2. The first factor within the sum was simplified using this assumption:

$$P(R_r^{m+1}| R_s^m, D_{d_m}^m \ldots, D_{d_1}^1, I) = P(R_r^{m+1}| R_s^m, I), \tag{22}$$

which in words states that past data are irrelevant for guessing whether the next observation starts a new subsequence or not, given that we know how many data belong to the current subsequence, and given that we *don't* yet know the value of the next observation.

According to this assumption, past observations help us guessing whether the next observation starts a new subsequence only if we can compare them with it. If we don't know it, past data can't help us. This assumption discards the possibility that particular values of past observations may trigger the start of a new sequence. The length of the current subsequence, R_s^m, is relevant in any case, because the next observation can only start a new subsequence or continue the previous one, so that either $r = 0$ or $r = s + 1$:

$$P(R_r^{m+1}| R_s^m, I) = h(s,m)\,\delta_{r,0} + [1-h(s,m)]\,\delta_{r,s+1} = \begin{cases} h(s,m) & \text{for } r = 0, \\ 1 - h(s,m) & \text{for } r = s + 1, \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

which simplifies our calculations: in the sum (22) when $r \geqslant 1$ only the term $s = r - 1$ survives. In our study we assume that $h(s,m)$ either is independent of s or increases with s, expressing our expectation that subsequences shouldn't be too long. This last option is reasonable if the participants were warned about the possibility of changepoints before the experiment.

3. The second factor within the sum is given by the Johnson-Dirichlet model, according to discussion of eq. (19).

4. The last factor within the sum is analogous to the joint probability for $R_r^{m+1}, D_{d_m}^m$, but calculated for the previous observation. In a recursive scheme it would have been computed in the preceding step.

In a recursive scheme we also need some initial values.

For $m = 0$ we have that $P(D_{d_1}^1| I)$ is given by the Johnson-Dirichlet model, and $P(R_0^1| I) = 1$ because the first observation is trivially a changepoint. Owing to the certainty of R_0^1, for $m = 1$ we have

$$P(R_r^2, D_{d_1}^1| I) = P(R_r^2, R_0^1, D_{d_1}^1| I) = P(R_r^2| R_0^1, I) \times P(D_{d_1}^1| I). \tag{24}$$

With the probabilities above, from $m = 2$ onward the recursion can be calculated with eqs (19)–(23).

4.2 The recursive algorithm

The recursive algorithm to calculate $P(D_{d_{m+1}}^{m+1} | D_{d_m}^m, \ldots, D_{d_1}^1, I)$ is shown in table 2. For $m = 0$ some initialization steps are necessary.

Table 1   Initial steps of predictive algorithm

1. Set $\Lambda = 0.01$, $\nu_d = 1/N$, $d \in \{1, \dots, N\}$
2. For $d_1 \in \{1, \dots, N\}$ calculate

$$A_1(d_1) := \mathrm{P}(D_{d_1}^1 \mid I) = \nu_{d_1}$$

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

$$C_2(r) := \mathrm{P}(R_r^2, D_{d_1}^1 \mid I) = A_1(d_1) \times \begin{cases} h(0,1) & \text{for } r = 0, \\ 1 - h(0,1) & \text{for } r = 1 \end{cases}$$

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

$$B_2(r, d_2) := \mathrm{P}(D_{d_2}^2 \mid R_r^2, D_{d_1}^1, I) = \frac{\Lambda \nu_{d_2} + r f_{d_2}}{\Lambda + r},$$

   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

$$A_2(d_2) := \mathrm{P}(D_{d_{12}}^2 \mid D_{d_1}^1, I) = \frac{\sum_{r=0}^1 B_2(r, d_2) \times C_2(r)}{A_1(d_1)}$$

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 $r \in \{0, \ldots, m\}$:

   1.1. Calculate

   $$C_{m+1}(r) := P(R_r^{m+1}, D_{d_m}^m \mid D_{d_{m-1}}^{m-1}, \ldots, D_{d_1}^1, I) =$$

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

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

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

   $$B_{m+1}(r, d_{m+1}) := P(D_{d_{m+1}}^{m+1} \mid R_r^{m+1}, D_{d_m}^m, \ldots, D_{d_{m+1-r}}^{m+1-r}, I) =$$

   $$\frac{\Lambda \nu_{d_{m+1}} + r f_{d_{m+1}}}{\Lambda + r},$$

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

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

   $$A_{m+1}(d_{m+1}) := P(D_{d_{m+1}}^{m+1} \mid D_{d_m}^m, \ldots, D_{d_1}^1, I) =$$

   $$\frac{\sum_{r=0}^m B_{m+1}(r, d_{m+1}) \times C_{m+1}(r)}{A_m(d_m)}$$

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

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

$$p(d_1, d_2 | \theta, I) = p(d_1 | \theta, I)\, p(d_2 | \theta, I) \tag{25}$$

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

$$p(d_2 | d_1, \theta, I) = p(d_2 | \theta, I), \tag{26}$$

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

$$p(\theta | d, I) = \frac{p(d | \theta, I)\, p(\theta | I)}{\sum_\theta p(d | \theta, I)\, p(\theta | I)}, \tag{27}$$

where $p(\theta | I)$ is the probability over a range of such models based only on knowledge I, and $p(d | \theta, I)$ 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

$$p(d_1, d_2 | \mu, I) = \sum_\theta p(d_1, d_2 | \theta, \mu, I)\, p(\theta | \mu, I)$$
$$\equiv \sum_\theta p(d_1 | \theta, \mu, I)\, p(d_2 | \theta, \mu, I)\, p(\theta | \mu, I), \tag{28}$$

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

$$p(d_2 | d_1, \mu, I) = \sum_\theta p(d_2 | \theta, \mu, I) \, p(\theta | d_1, \mu, I) \tag{29a}$$

with

$$p(\theta | d_1, \mu, I) = \frac{p(d_1 | \theta, \mu, I) \, p(\theta | \mu, I)}{\sum_\theta p(d_1 | \theta, \mu, I) \, p(\theta | \mu, I)}, \tag{29b}$$

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

$$p(\mu | d, I) = \frac{p(d | \theta, I) \, p(\mu | I)}{\sum_\mu p(d | \mu, I) \, p(\mu | I)} \tag{30a}$$

with

$$p(d | \mu, I) = \sum_\theta p(d | \theta, \mu, I) \, p(\theta | \mu, I). \tag{30b}$$

where $p(\mu | I)$ is the probability over a range of parametric models based only on knowledge I. The last expression (30b) is the *likelihood* of the model given the data, and we see that it's given by a mixture of independent models.

Equations (30) and (28) show that a parametric model is constructed as an uncertainty over independent models, and eq. (29) 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

$$p(d_1, d_2 | \chi, I) = \sum_\mu p(d_1, d_2 | \mu, \chi, I) \, p(\mu | \chi, I)$$

$$\equiv \sum_\mu \left\{ \sum_\theta \left[\prod_i p(d_i | \theta, \mu, \chi, I) \right] p(\theta | \mu, \chi, I) \right\} \times \tag{31}$$

$$p(\mu | \chi, I).$$

Learning takes place this way:

$$p(d_2 | d_1, \chi, I) = \sum_\mu \left[\sum_\theta p(d_2 | \theta, \mu, \chi, I) p(\theta | d_1, \mu, \chi, I) \right] p(\mu | d_1, \chi, I)$$

(32a)

with

$$p(\theta | d_1, \mu, \chi, I) = \frac{p(d_1 | \theta, \mu, \chi, I) p(\theta | \mu, \chi, I)}{\sum_\theta p(d_1 | \theta, \mu, \chi, I) p(\theta | \mu, \chi, I)}$$

$$= \frac{p(d_1 | \theta, \mu, \chi, I) p(\theta | \mu, \chi, I)}{p(d_1 | \mu, \chi, I)},$$

(32b)

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

(32c)

Note that the denominator of the update formula (32b) for θ is the updated probability (32c) 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 (31).

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:

$$\{\Theta\} := \bigcup_\mu \{\theta\}_\mu. \tag{33}$$

Now consider the predictive probability (31) for the hierarchic model:

$$p(d|\chi, I) = \sum_\mu \left[\sum_\theta p(d|\theta, \mu, \chi, I)\, p(\theta|\mu, \chi, I) \right] p(\mu|\chi, I). \tag{34}$$

If we decree that $p(\Theta|\mu, \chi, I) = 0$ 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 $p(d|\theta, \mu, \chi, I)$. The predictive probability above then becomes

$$\begin{aligned}
p(d|\chi, I) &= \sum_\Theta p(d|\Theta, \chi, I) \sum_\mu p(\Theta|\mu, \chi, I)\, p(\mu|\chi, I), \\
&= \sum_\Theta p(d|\Theta, \chi, I)\, p(\Theta|\chi, I),
\end{aligned} \tag{35}$$

where the last equality follows from the law of total probability. What's important in the last formula is that the probability $p(d|\Theta, \chi, I)$ 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 $p(d|\Theta, \chi, I)$. The probability $p(\Theta|\chi, I)$ 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.se as.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/detail s/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).