Bayesian Plinko

Study notes

Alex, Luca, Yasser, et al.

Draft of 29 May 2018 (first drafted 14 April 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 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 First study: exchangeable-model robot

2.1 The Bayesian robot

In the context of these notes and of the Plinko experiments (Filipowicz et al. 2014; 2016) we call 'model' any set of assumptions that allows us to assign a probability to a new observation, given a number of observations of a similar kind. Denote such assumptions by a proposition M – a proposition surely very difficult to express in writing. Denote the proposition 'The outcome of the ith observation is d' by D^i_d, with $d \in \{1, \dots N\}$. Then M allows us to give a numeric value to

$$\mathrm{P}(D^{m+1}_{d_{m+1}} | \, D^m_{d_m} \wedge \dots \wedge D^2_{d_2} \wedge D^1_{d_1} \wedge M), \tag{1}$$

We will abbreviate logical conjunction '$\wedge$' with a comma, for simplicity. Our statistical terminology and notation follow ISO standards (ISO 2009; 2006) otherwise.

We shall consider a robot who uses either of these two equivalent assumptions:

- the joint distribution for any number of observations is symmetric with respect to their order; that is, the order of the observations is irrelevant for inferential purposes;
- for inferential purposes, only the relative frequencies of past observations are relevant. Any additional data about past observation is irrelevant and can be discarded.

Distributions for different number of observations must of course be consistent with one another through marginalization.

The two equivalent assumptions are technically called *infinite exchangeability*. This notion was introduced by de Finetti (1930; 1937; Heath et al. 1976); it is described in detail in Bernardo et al. (2000 § 4.2).

Infinite exchangeability determines this form of the probability above:

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

where $\boldsymbol{q}$ is a normalized N-tuple of positive numbers: $\Delta := \{\boldsymbol{q} \in \mathbf{R}^N \mid q_i \geqslant 0, \sum_{i=1}^N q_i = 1\}$. This N-tuple can be thought of the relative, long-run frequencies of the possible outcomes[1], and $\mathrm{p}(\boldsymbol{q} | \, M) \, \mathrm{d}\boldsymbol{q}$ as their probability

[1] '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)

density. From this point of view it is as if the robot first assumes to know the long-run frequencies of the different 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, not being sure about the long-run frequencies, the robot assign to them the density $\mathrm{p}(\boldsymbol{q}|\, M)\,\mathrm{d}\boldsymbol{q}$ – which is determined by additional assumptions besides exchangeability.

As an explicit example, say with $N = 40$,

$$\mathrm{P}(D^1_{37}, D^2_6, D^3_{25}, D^4_{37}|\, M) = \int_{\Delta} q_6\, q_{25}\, {q_{37}}^2\, \mathrm{p}(\boldsymbol{q}|\, M)\,\mathrm{d}\boldsymbol{q}. \tag{3}$$

In the following we omit the integration domain Δ.

From Bayes's theorem we obtain the expression for the predictive probability (1) of an infinite exchangeable model:

$$\mathrm{P}(D^{m+1}_{d_{m+1}}|\, D^1_{d_1}, \dots, D^m_{d_m}, M) = \int q_{d_{m+1}} \mathrm{p}(\boldsymbol{q}|\, D^1_{d_1}, \dots, D^m_{d_m}, M)\,\mathrm{d}\boldsymbol{q}, \tag{4a}$$

$$\mathrm{p}(\boldsymbol{q}|\, D^1_{d_1}, \dots, D^m_{d_m}, M) = \frac{\left(\prod_{i=1}^{m} q_{d_i}\right) \mathrm{p}(\boldsymbol{q}|\, M)}{\int \left(\prod_{i=1}^{m} q'_{d_i}\right) \mathrm{p}(\boldsymbol{q}'|\, M)\,\mathrm{d}\boldsymbol{q}'}. \tag{4b}$$

Continuing our numeric example (3) this could be

$$\mathrm{P}(D^5_6|\, D^1_{37}, D^2_6, D^3_{25}, D^4_3, M) = \int q_6\ \mathrm{p}(\boldsymbol{q}|\, D^1_{37}, D^2_6, D^3_{25}, D^4_3, M)\,\mathrm{d}\boldsymbol{q}, \tag{5a}$$

$$\mathrm{p}(\boldsymbol{q}|\, D^1_{37}, D^2_6, D^3_{25}, D^4_3, M) = \frac{q_6\, q_{25}\, {q_{37}}^2\ \mathrm{p}(\boldsymbol{q}|\, M)}{\int q_6\, q_{25}\, {q_{37}}^2\ \mathrm{p}(\boldsymbol{q}'|\, M)\,\mathrm{d}\boldsymbol{q}'}. \tag{5b}$$

Formula (4) tell us how our robot would update its predictive probabilities at each new observation of a Plinko outcome.

2.2 General remarks on the robot's behaviour

The exchangeable-model formula (4) leads to some characteristic features of the robot's beliefs and of their evolution:

- The robot's predictions can be interpreted in several ways. One is this: the robot believes that there's 'something' constant in all trials;

loosely speaking, a 'constant mechanism'. Another, maybe preferable, is this: the robot has memory of past observations, but not of their order; any trends are therefore invisible to it.

- As data accumulate, the robot's probabilities for the next outcome approach the observed frequencies. Such approach happens independently of the form of the prior $\mathrm{p}(\boldsymbol{q}|\,M)\,\mathrm{d}\boldsymbol{q}$ – unless the latter is zero in peculiar regions of the integration domain – but the prior determines the celerity of the approach. A prior heavily peaked on a frequency $\boldsymbol{q}'$ will require a lot of data to move the predictions to a very different frequency $\boldsymbol{q}$.
- As data D accumulate, the updated density $\mathrm{p}(\boldsymbol{q}|\,D,M)\,\mathrm{d}\boldsymbol{q}$ will become more and more peaked at the N-tuple of observed frequencies.
- Suppose that we first have a long sequence of observations concentrating around frequencies $\boldsymbol{q}$ – say, a very long sequence of 1s in a row – and then a shift to other frequencies $\boldsymbol{q}'$ – say, suddenly 2s only appear. After the shift, the predictive 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.

2.3 Initial prior

The shape of the initial prior heavily determines the predictions in the first observations, so it must be chosen with care. The Plinko data tell us the initial predictive probabilities of the participants,

$$\mathrm{p}(D_k^1|\,M) \equiv \int q_k\ \mathrm{p}(\boldsymbol{q}|\,M)\,\mathrm{d}\boldsymbol{q}, \tag{6}$$

but not their prior $\mathrm{p}(\boldsymbol{q}|\,M)\,\mathrm{d}\boldsymbol{q}$.

As a first study we consider a *Johnson-Dirichlet* prior, proportional to a monomial $\prod_i {\boldsymbol{q}_i}^{x_i}$ for some values of x_i:

$$\mathrm{p}(\boldsymbol{q}|\,M_\mathrm{J}) = \frac{\Gamma(\Lambda)}{\prod_i \Gamma(\Lambda\,\nu_i)} \prod_{i=1}^{N} {\boldsymbol{q}_i}^{\Lambda\,\nu_i - 1}, \qquad \Lambda > 0, \nu \in \Delta. \tag{7}$$

This prior is determined by the additional assumption – call it M_J – that that the frequencies of other outcomes are irrelevant for predicting a particular one:

$$\mathrm{P}(D_k^{m+1}|\,N\boldsymbol{f}, M_\mathrm{J}) = \mathrm{P}(D_k^{m+1}|\,N f_k, M_\mathrm{J}) \qquad k \in \{1,\dots,N\}, \tag{8}$$

where $\boldsymbol{f}$ is the N-tuple of observed relative frequencies. This assumption is called 'sufficientness' (Johnson 1924; 1932; Good 1965 ch. 4; Zabell 1982; Jaynes 1996). This is a conjugate prior (DeGroot 2004 ch. 9; Diaconis et al. 1979) and it has two convenient properties: it updates to a density of the same mathematical form, and its corresponding predictive distribution can be calculated analytically using the formula

$$\int_{\Delta} \prod_{i=1}^{N} {q_i}^{x_i - 1} \,\mathrm{d}\boldsymbol{q} = \frac{\prod_i \Gamma(x_i)}{\Gamma\left(\sum_i x_i\right)}. \tag{9}$$

We obtain for the initial distribution:

$$\mathrm{P}(D_k^1 | \, M_{\mathrm{J}}) = \int q_k \; \mathrm{p}(\boldsymbol{q} | \, M_{\mathrm{J}}) \,\mathrm{d}\boldsymbol{q} = \nu_k, \tag{10}$$

and for the updated density:

$$\mathrm{p}(\boldsymbol{q} | \, D_{d_1}^1, \dots, D_{d_m}^m, M_{\mathrm{J}}) = \frac{\Gamma(\Lambda')}{\prod_i \Gamma(\Lambda' \nu_i')} \prod_{i=1}^{N} {\boldsymbol{q}_i}^{\Lambda' \nu_i' - 1}$$

$$\text{with} \quad \Lambda' = \Lambda + N, \quad \boldsymbol{\nu}' = \frac{\Lambda \boldsymbol{\nu} + N \boldsymbol{f}}{\Lambda + N}. \tag{11}$$

Formula (10) says that the Johnson-Dirichlet prior can produce any initial probabilities assigned by the participants, just by equalling the parameters $\boldsymbol{\nu}$ to them. The parameter Λ is left arbitrary. We can call it the *stubbornness* of the robot, Here's the reason.

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:

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

$$\text{and } \Lambda^{m+1} \coloneqq \Lambda^m + 1. \tag{12}$$

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 (11) says that the robot behaves as if it had already made Λ observations with outcome frequencies $\boldsymbol{\nu}$.

2.4 Examples: participant vs robot

Let's choose a participant, and use formula (10) to choose the ν 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 (11), 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.

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 § 2.2; 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

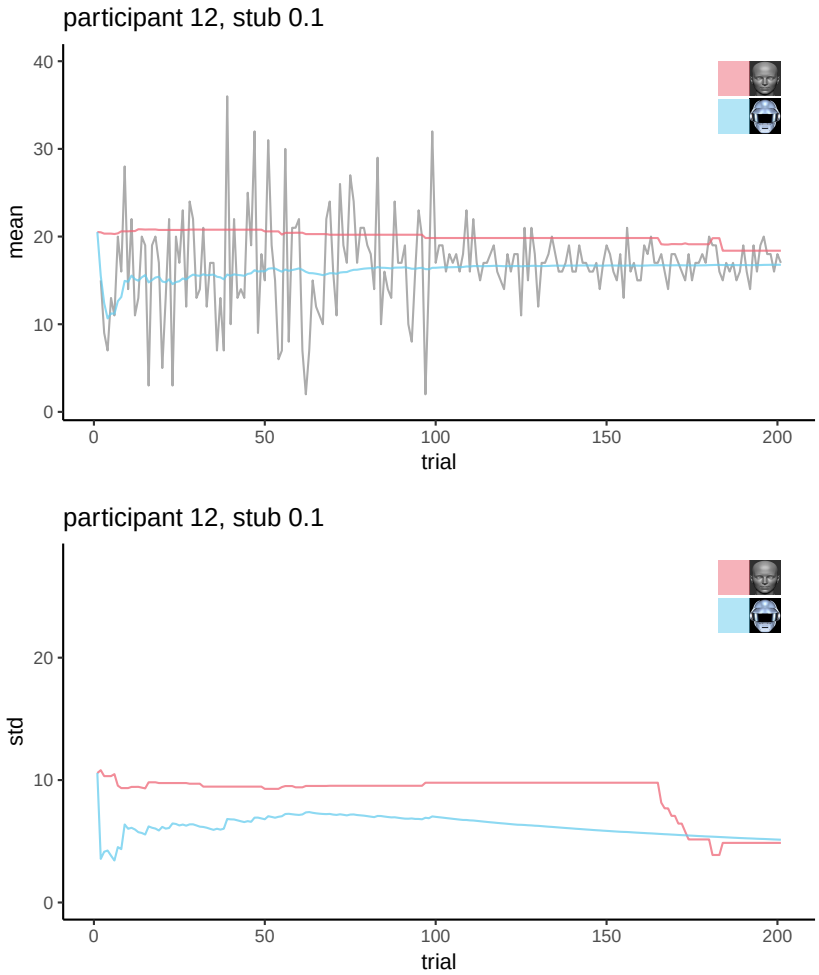

Figure 1 Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubborness $\Lambda = 0.1$

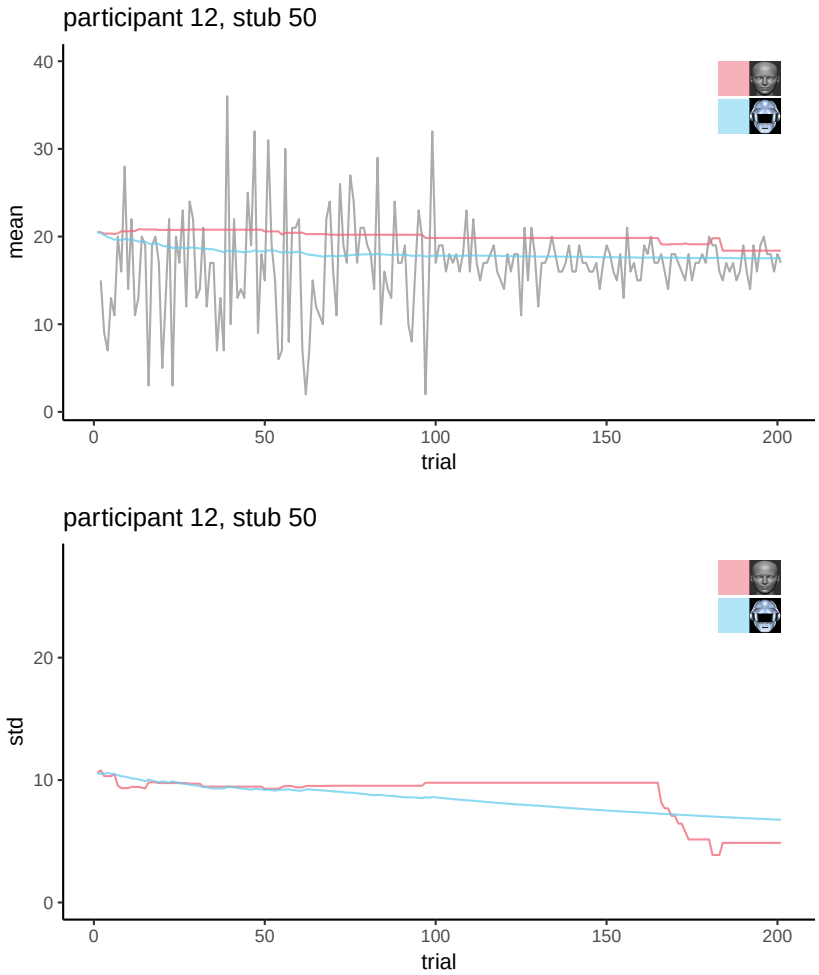

Figure 2 Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubborness $\Lambda = 50$

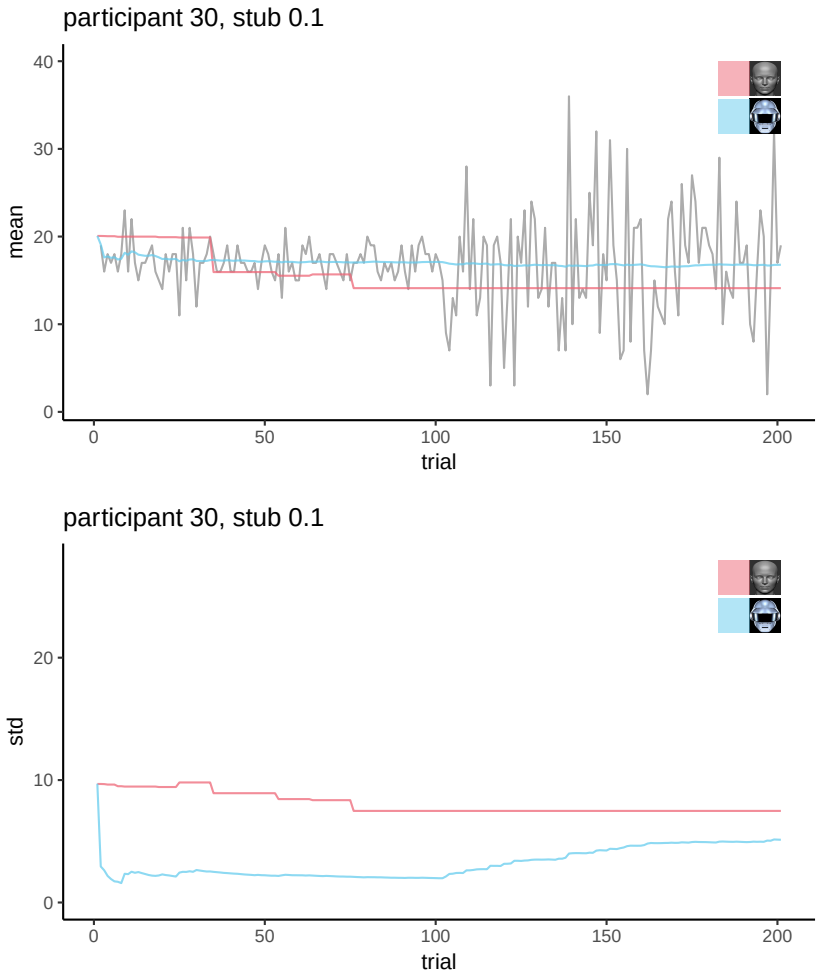

Figure 3 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubborness $\Lambda = 0.1$

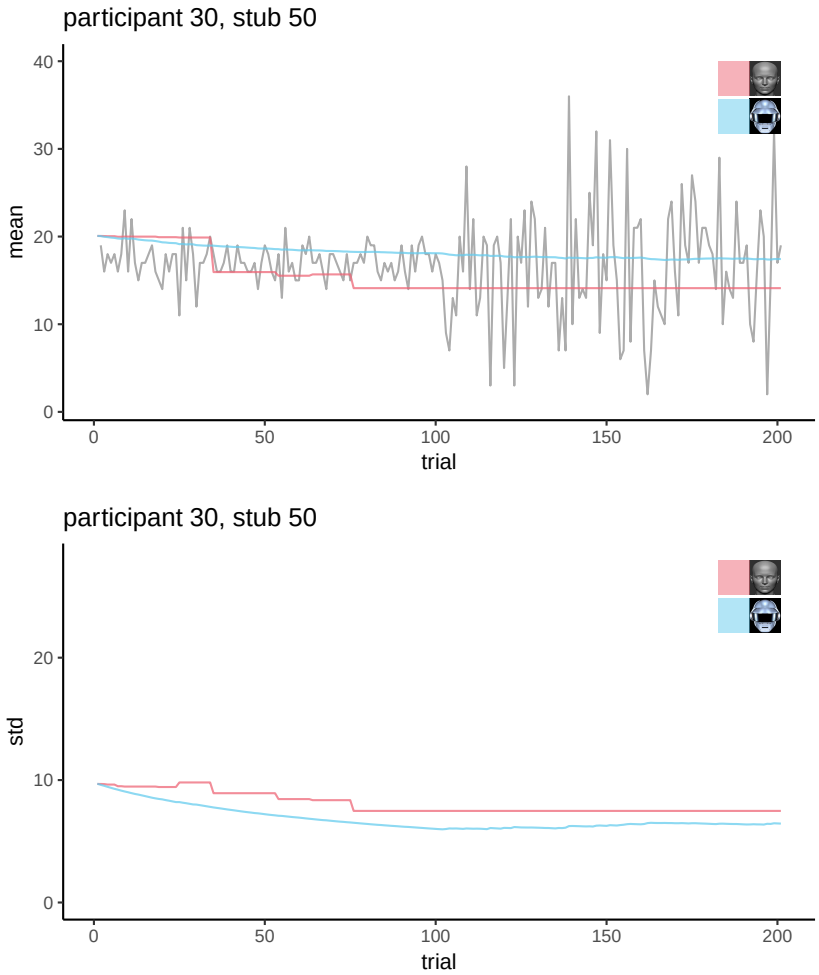

Figure 4 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubborness $\Lambda = 50$

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.

2.5 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.

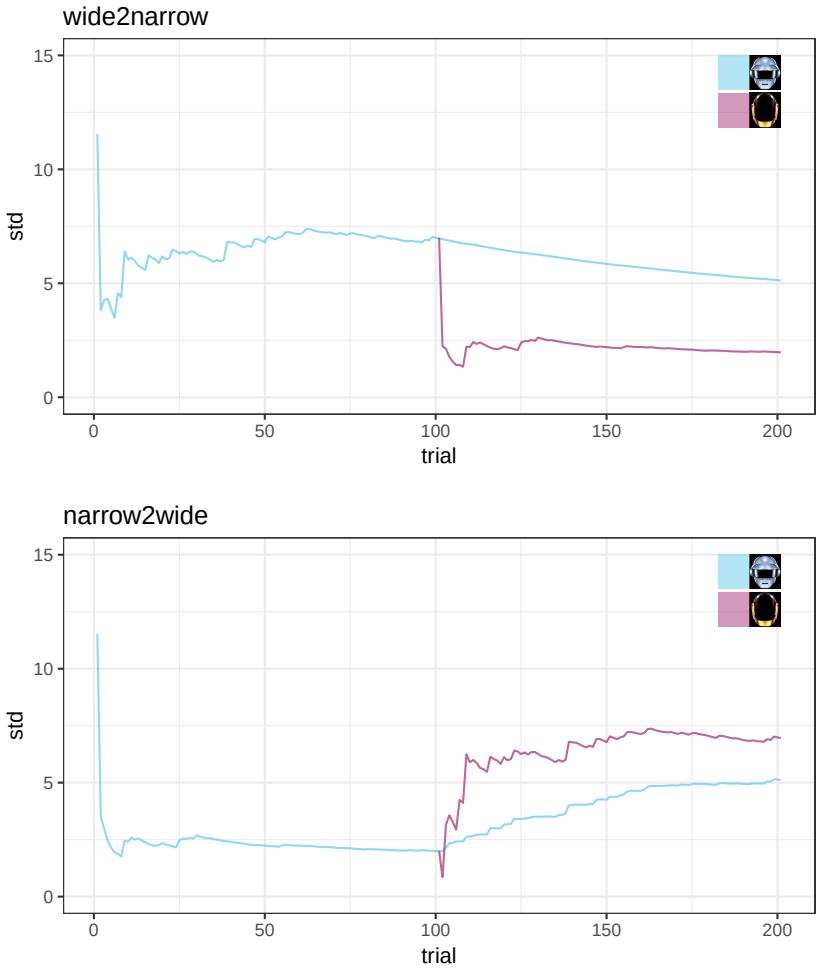

Figure 5 Adaptation speed of the standard deviation for a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change. The final relative entropy for the first robot relative to the second is 1.01 in the wide-to-narrow case, vs 0.260 in 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

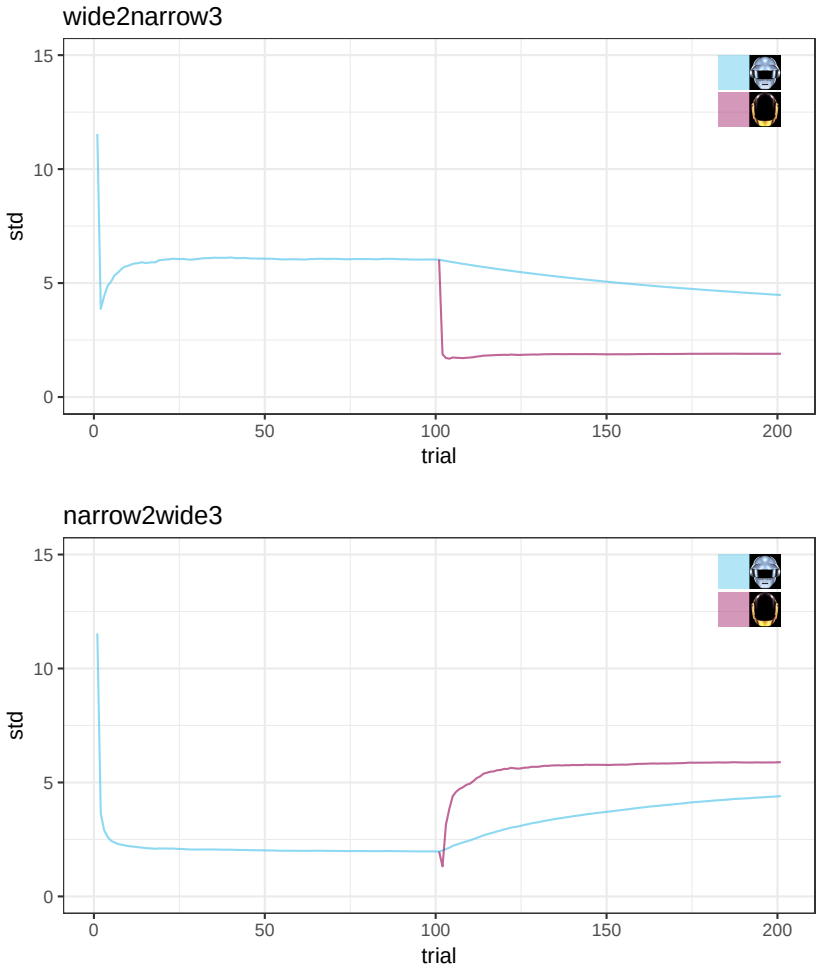

Figure 6 Adaptation speed of the standard deviation for a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change, averaged over 100 experiments. The normals have same mean 17 and standard deviations 6 and 1.9.

Figure 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.

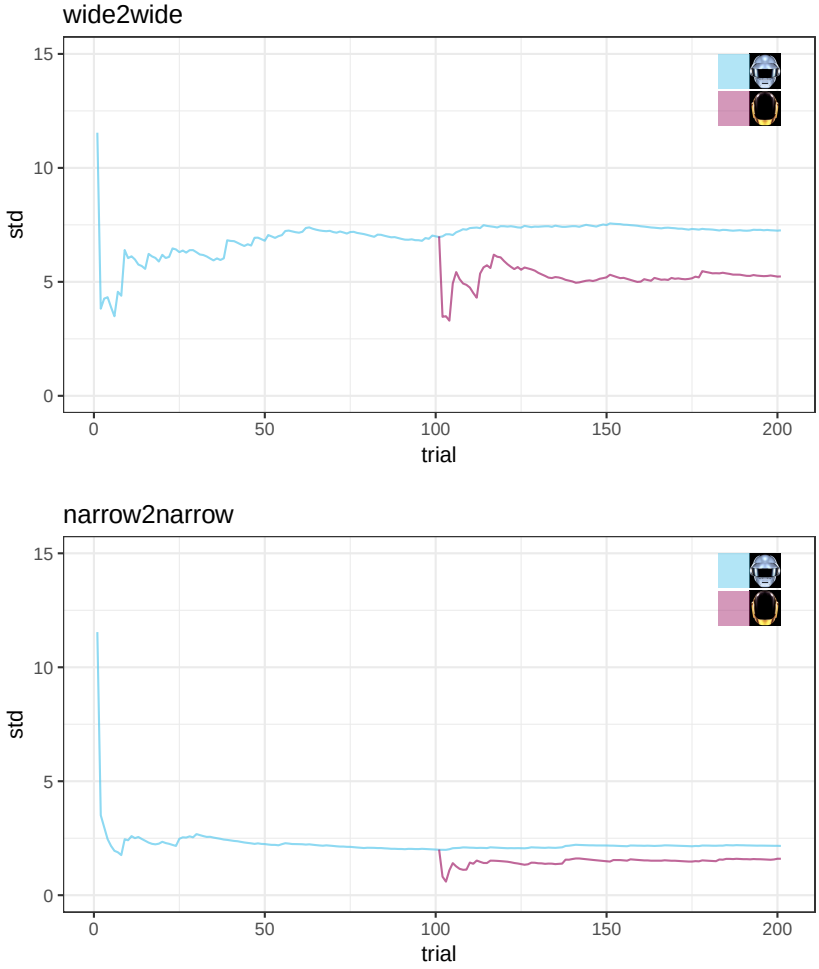

Figure 7 Adaptation speed of a robot after a change in the generation of the outcomes, compared with that of a robot that starts learning after the change.

3 Second study: exchangeable model with change-points

3.1 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 proportionally 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:

$$\begin{aligned}\mathrm{P}(\text{particular division into subsequences}|\,\text{data}, I) \propto \\ \textit{Johnson-Dirichlet} \begin{cases} \mathrm{P}(\text{data in 1st subsequence}|\,\text{particular division}, I) \\ \times\, \mathrm{P}(\text{data in 2nd subsequence}|\,\text{particular division}, I) \\ \times \cdots \end{cases} \\ \times\, \mathrm{P}(\text{particular division}|\, I), \end{aligned} \quad (13)$$

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

$$\mathrm{P}(\text{new observation}|\,\text{data}, I) \propto$$
$$\sum_{\text{all divisions}} \mathrm{P}(\text{new observation}|\,\text{division}, \text{data}, I) \times \mathrm{P}(\text{division}|\,\text{data}, I), \quad (14)$$

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 (13), 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^i_r. 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

$$\mathrm{P}(D^{m+1}_{d_{m+1}}|\, R^{m+1}_r, D^m_{d_m}, \ldots, D^1_{d_1}, I) =$$
$$\begin{cases} \mathrm{P}(D^{m+1}_{d_{m+1}}|\, R^{m+1}_r, D^m_{d_m}, \ldots, D^{m+1-r}_{d_{m+1-r}}, I) & \text{if } 1 \leqslant r \leqslant m, \\ \mathrm{P}(D^{m+1}_{d_{m+1}}|\, R^{m+1}_r, I) & \text{if } r = 0, \end{cases}$$
with probabilities given by the Johnson-Dirichlet model (4), (7), (12). (15)

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

$$\mathrm{P}(D^{m+1}_{d_{m+1}}|\, D^m_{d_m}, \ldots, D^1_{d_1}, I) =$$
$$\sum_{r=0}^{m} \mathrm{P}(D^{m+1}_{d_{m+1}}|\, R^{m+1}_r, D^m_{d_m}, \ldots, D^{m+1-r}_{d_{m+1-r}}, I) \times \mathrm{P}(R^{m+1}_r|\, D^m_{d_m}, \ldots, D^1_{d_1}, I). \quad (16)$$

The products in the sum contain the term given by (15) and the probability for R^{m+1}_r given the data.

This distribution for R_r^{m+1} represents the robot's guess of whether the new observation starts a new subsequence or whether it is part of the previous subsequence, and of how long that subsequence is. It can be calculated recursively by introducing R_s^m for the previous observation:

$$\mathrm{P}(R_r^{m+1} | D_{d_m}^m, \ldots, D_{d_1}^1, I) = \sum_{s=0}^{m-1} \mathrm{P}(R_r^{m+1} | R_s^m, I) \times \mathrm{P}(R_s^m | D_{d_{m-1}}^{m-1} \ldots, D_{d_1}^1, I). \tag{17}$$

Let's examine the two factors in the sum.

The first factor was simplified using this assumption:

$$\mathrm{P}(R_r^{m+1} | R_s^m, D_{d_m}^m \ldots, D_{d_1}^1, I) = \mathrm{P}(R_r^{m+1} | R_s^m, I), \tag{18}$$

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 $r = s + 1$:

$$\mathrm{P}(R_r^{m+1} | R_s^m, I) = h(s)\,\delta_{r,0} + [1 - h(s)]\,\delta_{r,s+1} = \begin{cases} h(s) & \text{for } r = 0, \\ 1 - h(s) & \text{for } r = s + 1, \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

which simplifies our calculations: in the sum (17) when $r \geqslant 1$ only the term $s = r - 1$ survives. In our study we assume that $h(s)$ 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.

The second factor in (17) is a distribution analogous to the one calculated there, but for the previous observation. In our recursive scheme it was already calculated in the previous observation and saved into memory.

3.2 The recursive algorithm

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

0.1. Set $m = 0$. Using the Johnson-Dirichlet model calculate $\mathrm{P}(D^{1}_{d_1} | I)$

0.2. Keep $B_1(0) \equiv 1$ for next steps

0.3. Observe d_1, set $m = 1$

1. For $0 \leqslant r \leqslant m$:

   1.1. Using the Johnson-Dirichlet model calculate

$$A(r) \coloneqq \mathrm{P}(D^{m+1}_{d_{m+1}} | R^{m+1}_{r}, D^{m}_{d_m}, \ldots, D^{m+1-r}_{d_{m+1-r}}, I)$$

   1.2. Calculate

$$B_{m+1}(r) \coloneqq \mathrm{P}(R^{m+1}_{r} | D^{m}_{d_m}, \ldots, D^{1}_{d_1}, I) = \begin{cases} \sum_{s=0}^{m-1} h(s) \times B_m(s) & \text{if } r = 0, \\ [1 - h(r-1)] \times B_m(r-1) & \text{if } r \geqslant 1 \end{cases}$$

2. Calculate

$$\mathrm{P}(D^{m+1}_{d_{m+1}} | D^{m}_{d_m}, \ldots, D^{1}_{d_1}, I) = \sum_{r=0}^{m} A(r) \times B_{m+1}(r)$$

3. Keep $\{B_{m+1}(r) \mid 0 \leqslant r \leqslant m\}$ for next steps

4. Observe d_{m+1}, increase m by 1, go to step 1

Table 1 Predictive algorithm

4 Notes on hierarchic models

4.1 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

$$\mathrm{p}(d_1, d_2 | \, \theta, I) = \mathrm{p}(d_1 | \, \theta, I)\, \mathrm{p}(d_2 | \, \theta, I) \tag{20}$$

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

$$\mathrm{p}(d_2 | \, d_1, \theta, I) = \mathrm{p}(d_2 | \, \theta, I), \tag{21}$$

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

$$\mathrm{p}(\theta | \, d, I) = \frac{\mathrm{p}(d | \, \theta, I)\, \mathrm{p}(\theta | \, I)}{\sum_\theta \mathrm{p}(d | \, \theta, I)\, \mathrm{p}(\theta | \, I)}, \tag{22}$$

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

$$\begin{aligned} \mathrm{p}(d_1, d_2 | \, \mu, I) &= \sum_\theta \mathrm{p}(d_1, d_2 | \, \theta, \mu, I)\, \mathrm{p}(\theta | \, \mu, I) \\ &\equiv \sum_\theta \mathrm{p}(d_1 | \, \theta, \mu, I)\, \mathrm{p}(d_2 | \, \theta, \mu, I)\, \mathrm{p}(\theta | \, \mu, I), \end{aligned} \tag{23}$$

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

$$\mathrm{p}(d_2|\,d_1, \mu, I) = \sum_\theta \mathrm{p}(d_2|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,d_1, \mu, I) \tag{24a}$$

with

$$\mathrm{p}(\theta|\,d_1, \mu, I) = \frac{\mathrm{p}(d_1|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,\mu, I)}{\sum_\theta \mathrm{p}(d_1|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,\mu, I)}, \tag{24b}$$

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

$$\mathrm{p}(\mu|\,d, I) = \frac{\mathrm{p}(d|\,\theta, I)\,\mathrm{p}(\mu|\,I)}{\sum_\mu \mathrm{p}(d|\,\mu, I)\,\mathrm{p}(\mu|\,I)} \tag{25a}$$

with

$$\mathrm{p}(d|\,\mu, I) = \sum_\theta \mathrm{p}(d|\,\theta, \mu, I)\,\mathrm{p}(\theta|\,\mu, I). \tag{25b}$$

where $\mathrm{p}(\mu|\,I)$ is the probability over a range of parametric models based only on knowledge I. The last expression (25b) is the *likelihood* of the model given the data, and we see that it's given by a mixture of independent models.

Equations (25) and (23) show that a parametric model is constructed as an uncertainty over independent models, and eq. (24) 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

$$\begin{aligned}\mathrm{p}(d_1, d_2|\,\chi, I) &= \sum_\mu \mathrm{p}(d_1, d_2|\,\mu, \chi, I)\,\mathrm{p}(\mu|\,\chi, I) \\ &\equiv \sum_\mu \Big\{ \sum_\theta \Big[\prod_i \mathrm{p}(d_i|\,\theta, \mu, \chi, I) \Big]\,\mathrm{p}(\theta|\,\mu, \chi, I) \Big\} \times \\ &\qquad\qquad \mathrm{p}(\mu|\,\chi, I).\end{aligned} \tag{26}$$

Learning takes place this way:

$$\mathrm{p}(d_2|\, d_1, \chi, I) = \sum_{\mu} \Big[\sum_{\theta} \mathrm{p}(d_2|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, d_1, \mu, \chi, I) \Big]\, \mathrm{p}(\mu|\, d_1, \chi, I) \tag{27a}$$

with

$$\begin{aligned} \mathrm{p}(\theta|\, d_1, \mu, \chi, I) &= \frac{\mathrm{p}(d_1|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, \mu, \chi, I)}{\sum_{\theta} \mathrm{p}(d_1|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, \mu, \chi, I)} \\ &= \frac{\mathrm{p}(d_1|\, \theta, \mu, \chi, I)\, \mathrm{p}(\theta|\, \mu, \chi, I)}{\mathrm{p}(d_1|\, \mu, \chi, I)}, \end{aligned} \tag{27b}$$

$$\mathrm{p}(\mu|\, d_1, \chi, I) = \frac{\mathrm{p}(d_1|\, \mu, \chi, I)\, \mathrm{p}(\mu|\, \chi, I)}{\sum_{\mu} \mathrm{p}(d_1|\, \mu, \chi, I)\, \mathrm{p}(\mu|\, \chi, I)}. \tag{27c}$$

Note that the denominator of the update formula (27b) for θ is the updated probability (27c) 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.

4.2 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 (26).

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:

$$\{\Theta\} \coloneqq \bigcup_\mu \{\theta\}_\mu. \tag{28}$$

Now consider the predictive probability (26) for the hierarchic model:

$$\mathrm{p}(d|\,\chi, I) = \sum_\mu \Big[\sum_\theta \mathrm{p}(d|\,\theta, \mu, \chi, I)\,\mathrm{p}(\theta|\,\mu, \chi, I)\Big]\,\mathrm{p}(\mu|\,\chi, I). \tag{29}$$

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

$$\begin{aligned} \mathrm{p}(d|\,\chi, I) &= \sum_\Theta \mathrm{p}(d|\,\Theta, \chi, I) \sum_\mu \mathrm{p}(\Theta|\,\mu, \chi, I)\,\mathrm{p}(\mu|\,\chi, I), \\ &= \sum_\Theta \mathrm{p}(d|\,\Theta, \chi, I)\,\mathrm{p}(\Theta|\,\chi, I), \end{aligned} \tag{30}$$

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.

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

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

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

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).

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).

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.

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.

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).