Bayesian Plinko

A. L. S. Filipowicz <***@***>
P.G.L. Porta Mana <pgl@portamana.org>

Draft of 15 April 2018 (first drafted 14 April 2018)

How does a Bayesian robot do at Plinko, using an infinitely exchangeable model and borrowing a human participant's prior?

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

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

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 an *infinitely exchangeable* model. This kind of models, introduced by de Finetti (1930; 1937; Heath et al. 1976) and described in detail in Bernardo et al. (2000 § 4.2), is determined by the following assumption of *infinite exchangeability*: 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. Distributions for different number of observations must of course be consistent with one another through marginalization. Infinite exchangeability may in turn be motivated by other specific assumptions, but the details of these are irrelevant for the mathematical form of this model.

Infinite exchangeability determines this form of the probability above:

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

where $\boldsymbol{q}$ is a normalized N-tuple of positive numbers: $\Delta \coloneqq \{\boldsymbol{q} \in \mathbf{R}^N \mid q_i \geqslant 0, \sum_{i=1}^N q_i = 1\}$. 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 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.

[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)

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:

- 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.

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 exploration 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

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

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.

What happens during the sequence of observations is this. Suppose that after some observations the predictive distribution is $\boldsymbol{\nu}$, and that the next outcome is k. Then the probability for slot k is updated to $\frac{\Lambda\nu_k+1}{\Lambda+1}$, and that for all other slots j to $\frac{\Lambda\nu_j}{\Lambda+1}$. 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 $\boldsymbol{\nu}$.

Some conclusions drawn from the formulae of this specific model:

- 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.
- 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.
- 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.

  ✠ L: to be continued

4 Methodological remarks

Bibliography

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

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

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, https://archive.org/details/johnsonslogic03johnuoft.

— (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.

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).