What is a probability model?

Luca
<piero.mana@ntnu.no>

Draft of 5 January 2019 (first drafted 21 October 2017)

***

> Where do probability models come from? To judge by the resounding silence over this question on the part of most statisticians, it seems highly embarrassing
>
> A. P. Dawid (1982 p. 220)

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

1 Motivation

The Bayesian literature on probabilistic modelling, including model comparison, hypothesis testing, and parameter estimation, often consider probability 'models' of this parametric form:

$$\mathrm{p}(\mathrm{data} \mid H) = \int \mathrm{p}(\mathrm{data} \mid \mathrm{parameters}, H)\, \mathrm{p}(\mathrm{parameters} \mid H)\, \mathrm{dparameters}, \tag{1}$$

where H represents the model. This is a particular case of an exchangeable model (Bernardo et al. 2000 ch. 4). Often discussed in this literature are Bayes factors and 'evidence' (Good 1985; MacKay 1992; Kass 1993; Kass et al. 1995), and model averaging (Draper 2005; Chatfield 1995; Draper 1995; Hoeting et al. 1999). In particular, sometimes there are difficulties in comparing the relative plausibilities of models or calculating their Bayes factors (Kass et al. 1995; O'Hagan 1995; Berger et al. 1996; De Santis et al. 1997; Berger et al. 1998). Intuitively, if we have several models H_1, H_2, the plausibility of each given some data is

$$\mathrm{p}(H_j \mid \mathrm{data}, I) = \frac{\mathrm{p}(\mathrm{data} \mid H_j, I)\, \mathrm{p}(H_j \mid I)}{\sum_j \mathrm{p}(\mathrm{data} \mid H_j, I)\, \mathrm{p}(H_j \mid I)}, \tag{2}$$

which involves the *evidence* $\mathrm{p}(\mathrm{data} \mid H, I)$, eq. (1). The Bayes factor between two models is just the ratio of their evidences. To calculate the evidence we solve the integral in eq. (1), and this requires specifying a distribution for the parameters, $\mathrm{p}(\mathrm{parameters} \mid H)$.

2 A tentative definition

The most general useful definition of a probability model seems to be this: given a set Y of possible data, and possibly a set X of conditional data, a *probability model* is a conjunction M of assumptions or hypotheses that allows us to assign a definite, numerical plausibility

$$\mathrm{p}(y_1, y_2, \dots \,|\, x_1, x_2, \dots, M) \tag{3}$$

for every meaningful combination of $y_i \in Y$ and $x_i \in X$. For the moment I consider finite sets.

This definition applies in particular to exchangeable models, where X is just a set of labels for the observations, which can take values in Y; to partially exchangeable models, where X is a set of labels for the exchangeable categories; and to models used in machine learning and neural nets. Possibly it also applies to time series. This definition also include functions or maps $f\colon X \to Y$ as special cases, when the plausibility is unity for a particular y only, dependent on x: $\mathrm{p}(y|\,x, M) = \delta[y, f(x)]$.

I'll use this definition to do some explorations in model comparison and parameter estimation, and especially in the notion of 'weight of evidence'.

We can make an important distinction between two kinds of model: *learning* models and *non-learning* models, which can also be called *extremal* for reasons explained later.

A learning model M is one that yields different plausibilities about some data $(y_1, y_2, \dots) =: \boldsymbol{y}$ conditional on $(x_1, x_2, \dots) =: \boldsymbol{x}$ if we condition on knowledge about other data $(y'_1, y'_2, \dots) =: \boldsymbol{y}'$, $(x'_1, x'_2, \dots) =: \boldsymbol{x}'$:

$$\mathrm{p}(\boldsymbol{y}|\,\boldsymbol{x},\ \boldsymbol{y}', \boldsymbol{x}',\ M) \neq \mathrm{p}(\boldsymbol{y}|\,\boldsymbol{x},\ M). \tag{4}$$

A non-learning model is one for which these plausibilities are not affected:

$$\mathrm{p}(\boldsymbol{y}|\,\boldsymbol{x},\ \boldsymbol{y}', \boldsymbol{x}',\ M) = \mathrm{p}(\boldsymbol{y}|\,\boldsymbol{x},\ M). \tag{5}$$

This means that

$$\mathrm{p}(\boldsymbol{y}|\,\boldsymbol{x},\ M) = \prod_i \mathrm{p}(y_i|\,x_i,\ M). \tag{6}$$

It is *non-learning* models that people usually, ultimately seek. For example, a trained and ready-to-use neural net is a non-learning model.

Parametric models are non-learning models; in fact, each value of the (possibly multi-dimensional) parameter defines a specific non-learning model.

Old text

When we ask about the probability of a 'model' given the data, we're asking if a given region of limit relative frequencies is more probable than another. This may not be what we want to ask, because one region as a whole can have higher probability than another region, and yet a particular frequency in the second region may be more probable than any single frequency in the first region.

I a way, our real goal is to guess the limit frequency, not guess a region in frequency space, which may be quite arbitrary and whose shape has nothing to do with our predictions.

A solution to this is to reparameterize every 'model' with a coordinate that has the same meaning across them, and then work with the union of these models, forgetting about their individualities.

But does it make sense to ask whether the limit frequency distribution belongs to parametric family rather than another? It is like asking whether the limit frequency belongs to a submanifold (for example a curve) rather than another in the simplex, in the case with finite number of outcomes. The limit frequency belongs to many submanifolds at once.

If we really want to ask that question we should first choose a probability distribution in the whole limit-frequency space, a metric, and then determine the induced probability on the submanifold.

This kind of question may be useful if we are asking about several *experiments*, not just one. In this case it may make sense to ask whether their different limit frequencies belong to some common submanifold.

'Model' often seems to be mistakenly identified with the specification of likelihoods only, as if the specification of the parameter prior were not part of the model. Compare Kass et al. (1995): 'Bayes factors require priors on the parameters appearing in the models that represent the competing hypotheses' (p. 773) 'the prior distributions $\pi(\theta_k | H_k)$ on the parameters of each model must be specified' (p. 781), 'Sensitivity analysis concerns distributional forms for models $\mathrm{pr}(\mathbf{D} | \theta_k, H_k)$ as well as priors', 'In choosing priors, just as in choosing models for data distributions, simplifications are often made' (p. 781). But see also 'the prediction rule is derived from the model H_k *(i.e., likelihood and prior)*' (p. 777, emphasis added).

The probability of hypotheses like those – concerning whole regions of limit-frequency space – cannot be computed.

The problem of 'model dimensionality' is also misplaced because we identify models with likelihoods only. In reality the dimensionality of a model is determined by the parameter prior. In fact, the very choice of likelihood can be interpreted as the choice of a particular prior from a 'non-parametric' point of view. (Compare Kass (1995), end of § 6.1.)

Also the idea of model *selection* can be dangerous, because we may be discarding the model than contains the frequency with highest likelihood.

The evidence is just an average of cross-validations (or splitting, see Kass § 6.5). Naive cross-validation is testing the wrong hypothesis.

3 Prediction and forecast

Some notation: We assume to have a possibly infinite set of observations, each of which can yield one of N outcomes, labelled by integers i. The proposition $O_i^{(a)}$ denotes that outcome i is observed at the ath observation. Such propositions also contain information about the time or place where the outcome was observed, so that from a proposition like $O_{i_2}^{(2)} \wedge O_{i_4}^{(4)}$ we can for example infer the time interval $t^{(4)} - t^{(2)}$ between observations number 4 and 2.

A statistical model is a set of assumptions M that jointly allow us to consistently assign numerical values to the probabilities

$$\mathrm{P}(O_{i_{n+1}}^{(a_{n+1})} | \, O_{i_1}^{(a_1)} \wedge \cdots \wedge O_{i_n}^{(a_n)} \wedge M), \tag{7}$$

for any legitimate n and any sets of observations $\{a_1, \dots, a_{n+1}\}$ and outcomes $\{i_1, \dots, i_{n+1}\}$; 'consistently' means that these assignments are properly related by operations like marginalization.

This definition is very general; in fact it amounts to say that a model is an assignment of the probabilities for all possible conjunctions of n outcomes, for all legitimate n.

it includes exchangeable models of various kinds, models for time series and forecasts.

Now I'd like to make a distinction between two main classes of statistical models: those that 'learn' and those what 'don't learn'.

This distinction is clear within the subclass of infinitely exchangeable models: for any such model the probability above has the form

$$\int \mathrm{p}(i_{n+1}|\,\theta, M)\,\mathrm{p}(\theta|\,i_1, \dots, i_n, M)\,\mathrm{d}\theta, \tag{8a}$$

$$\mathrm{p}(\theta|\,i_1, \dots, i_n, M) \propto \Bigl[\prod_{k=1}^{n} \mathrm{p}(i_k|\,\theta, M)\Bigr]\,\mathrm{p}(\theta|\,M), \tag{8b}$$

where the specific form of $\mathrm{p}(i|\,\theta, M)$ and $\mathrm{p}(\theta|\,M)$ are determined by the model. Within this subclass, models that don't learn are characterized by $\mathrm{p}(\theta|\,M) = \delta(\theta - \theta^*)$, so that the probability for an outcome does not depend on knowledge of other outcomes:

$$\mathrm{P}\bigl(O^{(a_{n+1})}_{i_{n+1}}|\,O^{(a_1)}_{i_1} \land \dots \land O^{(a_n)}_{i_n} \land M\bigr) = \mathrm{P}\bigl(O^{(a_{n+1})}_{i_{n+1}}|\,M\bigr) \equiv \mathrm{p}(i_{n+1}|\,\theta^*, M). \tag{9}$$

Such a model doesn't 'learn' because it makes all knowledge about other observations irrelevant for the prediction of each observation.

Among all statistical models for a particular set of

**models (e.g. exchangeability for which accumulation of data leads to stable probabilities, and models (e.g. Markov) for which this doesn't happen.

Thanks

. . . to Mari & Miri for continuous encouragement and affection, and to Buster Keaton and Saitama for filling life with awe and inspiration. To the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Bibliography

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

Berger, J. O., Pericchi, L. R. (1996): *The intrinsic Bayes factor for model selection and prediction*. J. Am. Stat. Assoc. **91**^{433}, 109–122.

— (1998): *Accurate and stable Bayesian model selection: the median intrinsic Bayes factor*. Sankhyā A **60**^{1}, 1–18. https://www2.stat.duke.edu/~berger/papers/medianibf.html.

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 **copasetal1995**.

Dawid, A. P. (1982): *Intersubjective statistical models*. In: Koch, Spizzichino (1982), 217–232.

De Santis, F., Spezzaferri, F. (1997): *Alternative Bayes factors for model selection*. Can. J. Stat. **25**^{4}, 503–515.

Draper, D. (1995): *Assessment and propagation of model uncertainty*. J. Roy. Stat. Soc. B **57**^{1}, 45–70. See also discussion and reply in **spiegelhalteretal1995**. https://classes.soe.ucsc.edu/ams206/Winter05/draper.pdf.

— (2005): *On the relationship between model uncertainty and inferential/predictive uncertainty*. https://users.soe.ucsc.edu/~draper/writings.html, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.9402. First written 1993.

Good, I. J. (1985): *Weight of evidence: a brief survey*. In: **bernardoetal1985**, 249–270. With discussion by H. Rubin, T. Seidenfeld, and reply.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999): *Bayesian model averaging: a tutorial*. Stat. Sci. **14**^{4}, 382–412. See also **clydeetal1999**.

Kass, R. E. (1993): *Bayes factors in practice*. The Statistician **42**^{5}, 551–560. http://ecologia.ib.usp.br/bie5782/lib/exe/fetch.php?media=bie5782:00_curso_avancado:uriarte:kass_statistician_1993_bayesfactors.pdf.

Kass, R. E., Raftery, A. E. (1995): *Bayes factors*. J. Am. Stat. Assoc. **90**^{430}, 773–795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf; https://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf.

Koch, G., Spizzichino, F., eds. (1982): *Exchangeability in Probability and Statistics*. (North-Holland, Amsterdam).

MacKay, D. J. C. (1992): *Bayesian interpolation*. Neural Comp. **4**^{3}, 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html.

O'Hagan, A. (1995): *Fractional Bayes factors for model comparison*. J. Roy. Stat. Soc. B **57**^{1}, 99–138.