The relation between cross-validation log-scores, weight of evidence, and Bayes factors

P.G.L. Porta Mana
Kavli Institute, Trondheim, Norway <piero.mana@ntnu.no>

Draft of 11 August 2019 (first drafted 8 August 2019)

***

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

1 ***

The probability calculus unequivocally tells us how our degree of belief in a hypothesis H_h given data D and background knowledge K, that is, $\mathrm{P}(H_h \mid D\, K)$, is related to our degree of belief in observing those data when we entertain that hypothesis as true, that is, $\mathrm{P}(D \mid H_h\, K)$:

$$\mathrm{P}(H_h \mid D\, K) = \frac{\mathrm{P}(D \mid H_h\, K)\, \mathrm{P}(H_h \mid K)}{\sum_h \mathrm{P}(D \mid H_h\, K)\, \mathrm{P}(H_h \mid K)}. \tag{1}$$

The term $\mathrm{P}(D \mid H_h\, K)$ is called the *likelihood* of the hypothesis given the data (Good 1950 § 6.1, p. 62). D, H_h, K denote propositions, which usually are about numeric quantities. I'll use the terms 'degree of belief', 'belief', and 'probability' as synonyms.

The post-data belief (1) is of course conditional on the set $\{H_h\}$ of mutually exclusive and exhaustive hypotheses under consideration, which is implicit in the knowledge K – in fact, the above formula is only valid if

$$\mathrm{P}\bigl(\bigvee_h H_h \mid K\bigr) = 1, \qquad \mathrm{P}(H_h \wedge H_{h'} \mid K) = 0 \quad \forall h \neq h'. \tag{2}$$

Only in extremely rare cases does the set of hypotheses $\{H_h\}$ encompass and reflect the extremely complex and fuzzy hypotheses that lie in the backs of our minds. The background knowledge K is therefore only a simplified picture of our actual knowledge. That's why K is called a *model*. 'A theory cannot duplicate nature, for if it did so in all respects, it would be isomorphic to nature itself and hence useless, a mere repetition of all the complexity which nature presents to us, that very complexity we frame theories to penetrate and set aside. If a theory were not simpler

than the phenomena it was designed to model, it would serve no purpose. Like a portrait, it can represent only a part of the subject it pictures. This part it exaggerates, if only because it leaves out the rest. Its simplicity is its virtue, provided the aspect it portrays be that which we wish to study' (**truesdelletal1980**).

We can combine our post-data beliefs in two hypotheses in different ways to form a quantitative idea of how much the data is giving evidence for one or the other. For example their ratio or the logarithm of their ratio, often called *weight of evidence* and *Bayes factor* (Good 1950 ch. 6; Osteyee et al. 1974 § 1.4; MacKay 1992; Kass et al. 1995; see also Jeffreys 1983 chs V, VI, A). 'It is historically interesting that the expression "weight of evidence", in its technical sense, anticipated the term "likelihood" by over forty years' (Osteyee et al. 1974 § 1.4.2 p. 12). I'm here speaking of the *comparison* of hypotheses, not of their *selection* – that is, of choosing one based on the data and discarding the others for future calculations. For such selection, probabilities aren't enough: we also need to specify a utility or cost function to calculate the expected gains of choosing one or another hypothesis.

The literature in probability and statistics has also employed various other ad-hoc measures to assess our beliefs in a set of hypotheses given some data. Here I consider one in particular, the *leave-one-out cross-validation log-score* (**krnjajicetal2014**; Krnjajić et al. 2011; Geisser et al. 1979; Vehtari et al. 2002; 2012; Piironen et al. 2017):

$$\frac{1}{d}\sum_{i=1}^{d} \ln \mathrm{P}(D_i \mid D_{-i}\, H_h\, K) \tag{3}$$

where every D_i is one datum in the data $D \equiv \bigwedge_i D_i$, and D_{-i} denotes the data with datum D_i excluded. The intuition behind this measure is more or less this: 'let's see what my belief in one datum should be, on average, once I've observed the other data, if I consider H_h as true'. 'On average' means considering such belief for every single datum in turn, and then taking the geometric mean, which is the arithmetic mean on a log scale. Krnjajić & Draper (**krnjajicetal2014**; 2011) compare this log-score with the deviance information criterion (There's an ambiguity in this definition, because we can ask: what's a 'datum' in the case of multi-dimensional observations? a single numerical value? or a multidimensional point? Different interpretations lead to different log-scores. We'll come back to this point later.)

This is a reasonable intuition, and the log-score (3) and post-data probability (1) often lead to qualitatively similar results in comparing two hypotheses. There are exceptions, though.

My point of view, which hinges on the logical foundations of the probability calculus (Pólya 1941; 1949; 1968; Cox 1946; Hailperin 1996; Jaynes 2003; Paris 2006; Snow 1998; Terenin et al. 2017), is that every intuitively built quantitative assessment of belief is either (1) an approximation of a formula that can be derived from the probability calculus, or (2) wrong.

I shall now show that the log-score above can be viewed as an approximation of the log-likelihood $\mathrm{P}(D \mid H_h\, K)$ of the post-data probability (1); or, if you like, that the post-data probability can be seen as a refined version of the log-score.

We can obviously write

$$\mathrm{P}(D \mid H\, K) \equiv \Big[\underbrace{\mathrm{P}(D \mid H\, K) \times \cdots \times \mathrm{P}(D \mid H\, K)}_{d \text{ times}}\Big]^{1/d} \tag{4}$$

where we have dropped the subscript $_h$ for simplicity. By the rules of probability we have

$$\mathrm{P}(D \mid H\, K) = \mathrm{P}(D_i \mid D_{-i}\, H_h\, K) \times \mathrm{P}(D_{-i} \mid H_h\, K) \tag{5}$$

and this holds no matter what specific $i \in \{1, \dots, d\}$ we choose (temporal ordering and similar matters are completely irrelevant in the formula above: it's a logical relation between propositions). So let's expand each of the d factors in the identity (4) using the product rule (5), but using a different datum in each of them. The result can be displayed thus:

$$\begin{aligned} \mathrm{P}(D \mid H\, K) \equiv \Big[& \mathrm{P}(D_1 \mid D_{-1}\, H\, K) \times \mathrm{P}(D_{-1} \mid H\, K) \times \\ & \mathrm{P}(D_2 \mid D_{-2}\, H\, K) \times \mathrm{P}(D_{-2} \mid H\, K) \times \\ & \qquad \dots \qquad \times \\ & \mathrm{P}(D_d \mid D_{-d}\, H\, K) \times \mathrm{P}(D_{-d} \mid H\, K) \Big]^{1/d}. \end{aligned} \tag{6}$$

$$\uparrow$$

log-score

Note that upon taking the logarithm of this expression, the d factors vertically aligned on the left add up to the log-score (3), as indicated.

Now note that the mathematical reshaping of $\mathrm{P}(D \mid H\,K)$ – that is, the root-product identity (4) and the expansion (6) – can be done for each of the remaining factors $\mathrm{P}(D_{-i} \mid H\,K)$, and so on recursively. Here is an explicit example for $d = 3$:

$$
\begin{aligned}
\mathrm{P}(D \mid H\,K) \equiv \\
&\Big\{ \mathrm{P}(D_1 \mid D_2\,D_3\,H\,K) \times \Big[\mathrm{P}(D_2 \mid D_3\,H\,K) \times \mathrm{P}(D_3 \mid H\,K) \times \\
&\qquad\qquad \mathrm{P}(D_3 \mid D_2\,H\,K) \times \mathrm{P}(D_2 \mid H\,K)\Big]^{1/2} \times \\
&\mathrm{P}(D_2 \mid D_1\,D_3\,H\,K) \times \Big[\mathrm{P}(D_1 \mid D_3\,H\,K) \times \mathrm{P}(D_3 \mid H\,K) \times \\
&\qquad\qquad \mathrm{P}(D_3 \mid D_1\,H\,K) \times \mathrm{P}(D_1 \mid H\,K)\Big]^{1/2} \times \\
&\mathrm{P}(D_3 \mid D_1\,D_2\,H\,K) \times \Big[\mathrm{P}(D_1 \mid D_2\,H\,K) \times \mathrm{P}(D_2 \mid H\,K) \times \\
&\qquad\qquad \mathrm{P}(D_2 \mid D_1\,H\,K) \times \mathrm{P}(D_1 \mid H\,K)\Big]^{1/2} \Big\}^{1/3}.
\end{aligned}
\tag{7}
$$

In this example, the logarithm of the three vertically aligned factors in the left column is, as already noted, the log-score (3). The logarithm of the six vertically aligned factors in the central column is an average of the log-scores calculated for the three distinct subsets of pairs of data $\{D_1\,D_2\}$, $\{D_1\,D_3\}$, $\{D_2\,D_3\}$. Likewise, the logarithm of the six factors vertically aligned on the right is the average of the log-scores for the three subsets of data singletons $\{D_1\}$, $\{D_2\}$, $\{D_3\}$.

In the general case with d data there are $\binom{d}{k}$ subsets with k data points. We therefore obtain

$$\begin{aligned}\ln \mathrm{P}(D \mid H\,K) \equiv{}& \frac{1}{d}\sum_{i=1}^{d} \ln \mathrm{P}(D_i \mid D_{-i}\,H\,K) + \\ & \frac{1}{d}\sum_{i\in\{1,\dots,d\}} \frac{1}{d-1} \sum_{j\in\{1,\dots,d\}}^{j\neq i} \ln \mathrm{P}(D_{-i,j} \mid D_{-i,-j}\,H\,K) + \cdots + \\ & \binom{d}{k}^{-1}\frac{1}{k} \sum_{k\text{-tuples}} \ln \mathrm{P}(D_{i_1} \mid \underbrace{D_{i_2}\cdots D_{i_k}}_{\text{order doesn't matter}}\,H\,K) + \cdots + \\ & \frac{1}{d}\sum_{i=1}^{d} \ln \mathrm{P}(D_i \mid H\,K), \\ \equiv{}& \sum_{k=1}^{d}\Bigl[k\,\tbinom{d}{k}\Bigr]^{-1} \sum_{k\text{-tuples}} \ln \mathrm{P}(D_{i_1} \mid \underbrace{D_{i_2}\cdots D_{i_k}}_{\text{order doesn't matter}}\,H\,K).\end{aligned} \tag{8}$$

The post-data log-probability for H will be equal to this expression, plus the pre-data log-probability for H, plus a term which is the same for all hypotheses.

I'd like to offer three ways of looking at the relation (8) between the log-likelihood and the log-score.

First, we can see the log-likelihood as a refinement and improvement of the log-score. The log-likelihood takes into account not only the log-score for the whole data, but also the log-scores for all possible subsets of data. Figuratively speaking it examines the relationship between hypothesis and data locally, globally, and on all intermediate scales.

The second point of view only holds for hypotheses $\hat{H}$ which make any observed data irrelevant:

$$\mathrm{P}(D \mid D'\,\hat{H}\,K) = \mathrm{P}(D \mid \hat{H}\,K) \qquad \text{if } D' \nRightarrow D, \tag{9}$$

or super-hypotheses H about such hypotheses, leading to exchangeable joint beliefs:

$$\mathrm{P}(DD' \mid H\,K) = \sum_h \mathrm{P}(D \mid \hat{H}_h\,H\,K)\,\mathrm{P}(D' \mid \hat{H}_h\,H\,K)\,\mathrm{P}(\hat{H}_h \mid H\,K) \qquad \text{if } D' \nRightarrow D. \tag{10}$$

In either case the log-score can be seen as an approximation of the log-likelihood; more precisely of the log-likelihood per datum:

$$\frac{1}{d}\sum_{i=1}^{d} \ln \mathrm{P}(D_i \mid D_{-i}\, H\, K) \approx \frac{1}{d} \ln \mathrm{P}(D \mid H\, K). \tag{11}$$

This is in fact an exact equality if property (9) holds for H.

which lead to exchangeable beliefs about the data

Second, we can see This approximation is only valid

This approximation is reasonable if the amount of data is large with respect to the dimension of the space of a single datum, because *** (ref to geisser, stone, gelfandetal)

*** problems calculation with time-relevant hypotheses

'we cannot give a universal rule for them beyond the common-sense one, that if anybody does not know what his suggested value is, or whether there is one, he does not know what question he is asking and consequently does not know what his answer means' (Jeffreys 1983 § 3.1, p. 124).

*** with similar procedure we can included all k-fold scores.

**

Bibliography

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

Cox, R. T. (1946): *Probability, frequency, and reasonable expectation*. Am. J. Phys. **14**[1], 1–13. http://jimbeck.caltech.edu/summerlectures/references/ProbabilityFrequencyReasonableExpectation.pdf, https://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/Cox_1946.pdf.

Draper, D., Krnjajić, M. (2014): *Bayesian model comparison: log scores and* DIC. Stat. Probab. Lett. **88**, 9–14.

Geisser, S., Eddy, W. F. (1979): *A predictive approach to model selection*. J. Am. Stat. Assoc. **74**[365], 153–160.

Good, I. J. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

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

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

Jeffreys, H. (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.

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.

Krnjajić, M., Draper, D. (2011): *Bayesian model specification: some problems related to model choice and calibration*. http://hdl.handle.net/10379/3804.

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

Osteyee, D. B., Good, I. J. (1974): *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection*. (Springer, Berlin).

Paris, J. B. (2006): *The Uncertain Reasoner's Companion: A Mathematical Perspective*, reprint. (Cambridge University Press, Cambridge). See also Snow (1998).

Piironen, J., Vehtari, A. (2017): *Comparison of Bayesian predictive methods for model selection*. Stat. Comput. **27**[3], 711–735.

Pólya, G. (1941): *Heuristic reasoning and the theory of probability*. Am. Math. Monthly **48**[7], 450–465. First written in French 1939.

— (1949): *Preliminary remarks on a logic of plausible inference*. Dialectica **3**[1–2], 28–35.

— (1968): *Mathematics and Plausible Reasoning: Vol. II: Patterns of Plausible Inference*, 2nd ed. (Princeton University Press, Princeton). First publ. 1954.

Snow, P. (1998): *On the correctness and reasonableness of Cox's theorem for finite domains*. Comput. Intell. **14**[3], 452–459.

Terenin, A., Draper, D. (2017): *Cox's theorem and the Jaynesian interpretation of probability*. arXiv:1507.06597. First publ. 2015.

Vehtari, A., Lampinen, J. (2002): *Bayesian model assessment and comparison using cross-validation predictive densities*. Neural Comp. **14**[10], 2439–2468.

Vehtari, A., Ojanen, J. (2012): *A survey of Bayesian predictive methods for model assessment, selection and comparison*. Statist. Surv. **6**, 142–228.