Open Science Framework DOI:10.31219/osf.io/k8mj3

Models and hypotheses, log-likelihoods, and cross-validation log-scores

P.G.L. Porta Mana
Kavli Institute, Trondheim <pgl@portamana.org>

18 August 2019; updated 17 May 2020

It is shown that the log-likelihood of a hypothesis or model given some data is equal to an average of all leave-one-out cross-validation log-scores that can be calculated from all subsets of the data. This relation can be generalized to any k-fold cross-validation log-scores.

1 Log-likelihoods and cross-validation log-scores

What is the probability of hypothesis H, given data D in some specific context I? It is $\mathrm{P}(H \mid D \land I)$. The probability-calculus gives a straight-forward relation between this probability and the probability of the data given the hypothesis, $\mathrm{P}(D \mid H \land I)$ – also called the *likelihood* for the hypothesis given the data[1]:

$$\mathrm{P}(H \mid D \land I) = \frac{\mathrm{P}(D \mid H \land I)\, \mathrm{P}(H \mid I)}{\mathrm{P}(D \mid I)} . \tag{1}$$

If we have hypotheses $\{H_h\}$ that are mutually exclusive on context I, and if their probabilities $\{\mathrm{P}(H_h \mid I)\}$ are all equal, then their likelihoods decide which among them is the most probable given the data, owing to the proportionality in the equation above.

Note that we are speaking about our degrees of belief in the hypotheses, not about the *choice* of one among them. Such a choice would require also a utility function, and the chosen hypothesis would not need to be the most probable one. But in fact we rarely have to 'choose' among hypotheses. We usually have to make a decision that can somewhat be interpreted *as if* we fully believed in a hypothesis. A clinician, for example, may give a patient some treatment for a disease and yet believe that the patient is healthy. Simply because the treatment will not harm the patient if he is healthy, and will cure him in the improbable case he is not. Likewise, an astrophysicist may use Newton's equations to

[1] Good 1950 § 6.1 p. 62.

This document is designed for screen reading and two-up printing on A4 or Letter paper

find the motion of a celestial object and yet firmly believe in the correctness of Einstein's equations. Simply because the approximate answer of the former is faster to compute and enough precise for the problem considered. Beliefs and decisions are different things.

I assume that the problem of *model comparison* often discussed in the literature ***

Despite the clear relation of equation (3a), the literature in probability and statistics employs and debates other ad-hoc measures to find which hypothesis should have our highest belief given the data.

Here I consider one measure in particular: the *leave-one-out cross-validation log-score*[1], which I'll just call 'log-score' for brevity

quantify how the data relate to the hypotheses – or even to select one hypothesis for further use, discarding the others[2]. Here I consider one measure in particular: the *leave-one-out cross-validation log-score*[2], which I'll just call 'log-score' for brevity:

$$\frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid D_{-i}\, H_h\, I) \tag{2}$$

where every D_i is one datum in the data $D \equiv \bigwedge_{i=1}^{d} D_i$, and D_{-i} denotes the data with datum D_i excluded. The intuition behind this score can be colloquially expressed thus: 'let's see what my belief in one datum would 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 of the resulting beliefs. Other variants of this score use more general partitions of the data into two disjoint subsets[2].

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

[2] Bernardo & Smith 2000 §§ 3.4, 6.1.6 gives the clearest motivation and explanation; see also Stone 1977; Geisser & Eddy 1979; Vehtari & Ojanen 2012; Vehtari & Lampinen 2002; Krnjajić & Draper 2011; 2014; Gelman et al. 2014; Gronau & Wagenmakers 2019; Chandramouli et al. 2019.

$\mathrm{P}(D \mid H_h\, I)$:

$$\mathrm{P}(H_h \mid D\, I) = \frac{\mathrm{P}(D \mid H_h\, I)\, \mathrm{P}(H_h \mid I)}{\mathrm{P}(D \mid I)} \tag{3a}$$

$$= \frac{\mathrm{P}(D \mid H_h\, I)\, \mathrm{P}(H_h \mid I)}{\sum_{h'} \mathrm{P}(D \mid H_{h'}\, I)\, \mathrm{P}(H_{h'} \mid I)}. \tag{3b}$$

D, H_h, I denote propositions, which are usually about numeric quantities. I use the terms 'degree of belief', 'belief', and 'probability' as synonyms. By 'hypothesis' I mean either a scientific (physical, biological, etc.) hypothesis – a state or development of things capable of experimental verification, at least in a thought experiment – or more generally some proposition, often not precisely specified, which leads to quantitatively specific distributions of beliefs for any contemplated data set. In the latter case we often call H_h a '(probabilistic) model' rather than a 'hypothesis'.

Expression (3b) assumes that we have a set $\{H_h\}$ of mutually exclusive and exhaustive hypotheses under consideration, which is implicit in our knowledge I. In fact it's only valid if

$$\mathrm{P}\bigl(\textstyle\bigvee_h H_h \mid I\bigr) = 1, \qquad \mathrm{P}(H_h \land H_{h'} \mid I) = 0 \quad \text{if } h \neq h'. \tag{4}$$

Only rarely does the set of hypotheses $\{H_h\}$ encompass and reflect the extremely complex and fuzzy hypotheses lying in the backs of our minds. They're simplified pictures. That's also why they're called 'models'.

Expression (3a) is universally valid instead, but it's rarely possible to quantify its denominator $\mathrm{P}(D \mid I)$ unless we simplify our inferential problem by introducing a possibly unrealistic exhaustive set of hypotheses, thus falling back to (3b). We can bypass this problem if we are content with comparing our beliefs about any two hypotheses through their ratio, so that the term $\mathrm{P}(D \mid I)$ cancels out. See Jaynes's[3] insightful remarks about such binary comparisons, and also Good's[4].

The term $\mathrm{P}(D \mid H_h\, I)$ in eq. (3) is called the *likelihood* of the hypothesis given the data[5]. Its logarithm is surprisingly called log-likelihood:

$$\log \mathrm{P}(D \mid H_h\, I), \tag{5}$$

where the logarithm can be taken in an arbitrary basis (Turing, Good[6], Jaynes[7] recommend base $10^{1/10}$, leading to a measurement in decibels; see the cited works for the practical advantages of such choice).

The ratio of the likelihoods of two hypotheses, called *relative Bayes factor*, or its logarithm, the *relative weight of evidence*,[8] are often used to quantify how much the data favour our belief in one versus the other hypothesis (that is, assuming at least momentarily that they be exhaustive). 'It is historically interesting that the expression "weight of evidence", in its technical sense, anticipated the term "likelihood" by over forty years'[9].

Recent literature[10] seems to exclusively deal with *relative* Bayes factors. I'd like to recall, lest it fades from the memory, the definition of the non-relative Bayes factor for a hypothesis H_h provided by data D:[11]

$$\frac{\mathrm{P}(D \mid H_h\, I)}{\mathrm{P}(D \mid \neg H_h\, I)} \equiv \frac{\mathrm{O}(H_h \mid D\, I)}{\mathrm{O}(H_h \mid I)} = \frac{\mathrm{P}(D \mid H_h\, I)\,[1-\mathrm{P}(H_h \mid I)]}{\sum_{h'}^{h'\neq h} \mathrm{P}(D \mid H_{h'}\, I)\,\mathrm{P}(H_{h'} \mid I)}, \tag{6}$$

where the *odds* O is defined as $\mathrm{O} \coloneqq \mathrm{P}/(1-\mathrm{P})$. Looking at the expression on the right, which can be derived from the probability rules, it's clear that the Bayes factor for a hypothesis involves the likelihoods of *all* other hypotheses as well as their pre-data probabilities. This quantity and its logarithm, the (non-relative) weight of evidence, have important properties which relative Bayes factors and relative weights of evidence don't enjoy. For example, the

[3] Jaynes 2003 §§ 4.3–4.4. [4] Good 1950 § 6.3–6.6. [5] Good 1950 § 6.1 p. 62. [6] e.g. Good 1985; 1950; 1969. [7] Jaynes 2003 § 4.2. [8] Good 1950 ch. 6; 1975; 1981; 1985, and many other works in Good 1983; Osteyee & Good 1974 § 1.4; MacKay 1992; Kass & Raftery 1995; see also Jeffreys 1983 chs V, VI, A. [9] Osteyee & Good 1974 § 1.4.2 p. 12. [10] for example Kass & Raftery 1995. [11] Good 1981 § 2.

expected weight of evidence for a correct hypothesis is always positive, and for a wrong hypotheses always negative[12]. See Jaynes[13] for further discussion and a numeric example.

The literature in probability and statistics has also employed and debated other ad-hoc measures to quantify how the data relate to the hypotheses – or even to select one hypothesis for further use, discarding the others[14]. Here I consider one measure in particular: the *leave-one-out cross-validation log-score*[14], which I'll just call 'log-score' for brevity:

$$\frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid D_{-i}\, H_h\, I) \tag{7}$$

where every D_i is one datum in the data $D \equiv \bigwedge_{i=1}^{d} D_i$, and D_{-i} denotes the data with datum D_i excluded. The intuition behind this score can be colloquially expressed thus: 'let's see what my belief in one datum would 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 of the resulting beliefs. Other variants of this score use more general partitions of the data into two disjoint subsets[14].

My purpose is to show an exact relation between the log-likelihood (5) and the leave-one-out cross-validation log-score (7). This relation doesn't seem to appear in the literature, and I find it very intriguing because it portrays the log-likelihood as a sort of full-scale use of the log-score: it says that *the log-likelihood is the sum of all averaged log-scores that can be formed from all data subsets*. The relation can be extended to more general cross-validation log-scores, and it can be of interest for the debate about the soundness of log-scores in deciding among hypotheses.

[12] Good 1950 § 6.7. [13] Jaynes 2003 §§ 4.3–4.4. [14] Bernardo & Smith 2000 §§ 3.4, 6.1.6 gives the clearest motivation and explanation; see also Stone 1977; Geisser & Eddy 1979; Vehtari & Ojanen 2012; Vehtari & Lampinen 2002; Krnjajić & Draper 2011; 2014; Gelman et al. 2014; Gronau & Wagenmakers 2019; Chandramouli et al. 2019.

2 A relation between log-likelihood and log-score

We can obviously write the likelihood as the dth root of its dth power:

$$\mathrm{P}(D \mid H\, I) \equiv \bigl[\underbrace{\mathrm{P}(D \mid H\, I) \times \dots \times \mathrm{P}(D \mid H\, I)}_{d \text{ times}}\bigr]^{1/d} \tag{8}$$

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

$$\mathrm{P}(D \mid H\, I) = \mathrm{P}(D_i \mid D_{-i}\, H_h\, I) \times \mathrm{P}(D_{-i} \mid H_h\, I) \tag{9}$$

no matter which 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 (8) using the product rule (9), using a different i for each of them. The result can be thus displayed:

$$\begin{aligned}\mathrm{P}(D \mid H\, I) \equiv \bigl[\, & \mathrm{P}(D_1 \mid D_{-1}\, H\, I) \times \mathrm{P}(D_{-1} \mid H\, I) \times \\ & \mathrm{P}(D_2 \mid D_{-2}\, H\, I) \times \mathrm{P}(D_{-2} \mid H\, I) \times \\ & \qquad\qquad \dots \qquad\qquad\qquad \times \\ & \underset{\substack{\uparrow \\ \text{this column leads to the log-score}}}{\mathrm{P}(D_d \mid D_{-d}\, H\, I)} \times \mathrm{P}(D_{-d} \mid H\, I)\,\bigr]^{1/d}.\end{aligned} \tag{10}$$

Upon taking the logarithm of this expression, the d factors vertically aligned on the left add up to the log-score (7), as indicated. But the mathematical reshaping we just did for $\mathrm{P}(D \mid H\, I)$ – that is, the root-product identity (8) and the expansion (10) – can be done for each of the remaining factors $\mathrm{P}(D_{-i} \mid H\, I)$ vertically aligned on the right in the expression above; and so on recursively. Here is an explicit example for

$d = 3$:

$$
\begin{aligned}
&\mathrm{P}(D \mid H\, I) \equiv \\
&\qquad \Bigl\{ \mathrm{P}(D_1 \mid D_2\, D_3\, H\, I) \times \bigl[\mathrm{P}(D_2 \mid D_3\, H\, I) \times \mathrm{P}(D_3 \mid H\, I) \times \\
&\qquad\qquad\qquad \mathrm{P}(D_3 \mid D_2\, H\, I) \times \mathrm{P}(D_2 \mid H\, I)\bigr]^{1/2} \times \\
&\qquad \mathrm{P}(D_2 \mid D_1\, D_3\, H\, I) \times \bigl[\mathrm{P}(D_1 \mid D_3\, H\, I) \times \mathrm{P}(D_3 \mid H\, I) \times \\
&\qquad\qquad\qquad \mathrm{P}(D_3 \mid D_1\, H\, I) \times \mathrm{P}(D_1 \mid H\, I)\bigr]^{1/2} \times \\
&\qquad \mathrm{P}(D_3 \mid D_1\, D_2\, H\, I) \times \bigl[\mathrm{P}(D_1 \mid D_2\, H\, I) \times \mathrm{P}(D_2 \mid H\, I) \times \\
&\qquad\qquad\qquad \mathrm{P}(D_2 \mid D_1\, H\, I) \times \mathrm{P}(D_1 \mid H\, I)\bigr]^{1/2} \Bigr\}^{1/3} . \qquad (11)
\end{aligned}
$$

In this example the logarithm of the three vertically aligned factors in the left column is, as already noted, the log-score (7). 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

$$
\begin{aligned}
\log \mathrm{P}(D \mid H\, I) \equiv{}& \frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid D_{-i}\, H\, I) + \\
& \frac{1}{d}\sum_{i\in\{1,\dots,d\}} \frac{1}{d-1}\sum_{j\in\{1,\dots,d\}}^{j\neq i} \log \mathrm{P}(D_{-i,j} \mid D_{-i,-j}\, H\, I) + \\
& \binom{d}{d-2}^{-1} \sum_{i,j\in\{1,\dots,d\}}^{i<j} \frac{1}{d-2} \sum_{k\in\{1,\dots,d\}}^{k\neq i,j} \log \mathrm{P}(D_{-i,-j,k} \mid D_{-i,-j,-k}\, H\, I) + \\
& \cdots + \\
& \binom{d}{2}^{-1} \sum_{i,j\in\{1,\dots,d\}}^{i<j} \frac{1}{2}\bigl[\log \mathrm{P}(D_i \mid D_j\, H\, I) + \log \mathrm{P}(D_j \mid D_i\, H\, I)\bigr] + \\
& \frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid H\, I),
\end{aligned}
\tag{12}
$$

which can be compactly written

$$
\boxed{\log \mathrm{P}(D \mid H\, I) \equiv \sum_{k=1}^{d} \binom{d}{k}^{-1} \sum_{\substack{\text{ordered}\\ k\text{-tuples}}} \frac{1}{k} \sum_{\substack{\text{cyclic}\\ \text{permutations}}} \log \mathrm{P}(D_{i_1} \mid D_{i_2} \cdots D_{i_k}\, H\, I).}
\tag{13}
$$

That is, *the log-likelihood is the sum of all averaged log-scores that can be formed from all (non-empty) data subsets with k elements*, the average for log-scores over k data being taken over the $\binom{d}{k}$ subsets having the same cardinality k.

There's also an equivalent form with a slightly different cross-validating interpretation: We take each datum D_j in turn and calculate our log-belief in it conditional on all possible subsets of remaining data, from the empty subset with no data (term $k = 0$), to the only subset D_{-j} with all data except D_j (term $k = d - 1$). These log-beliefs are averaged over the $\binom{d-1}{k}$ subsets having the same cardinality k. The result can be expressed as

$$\boxed{\log \mathrm{P}(D \mid H\, I) \equiv \frac{1}{d} \sum_{j=1}^{d} \sum_{k=0}^{d-1} \binom{d-1}{k}^{-1} \sum_{\substack{\text{ordered}\\ k\text{-tuples,}\\ j \text{ excluded}}} \log \mathrm{P}(D_j \mid D_{i_1} \cdots D_{i_k}\, H\, I).} \tag{14}$$

3 Brief discussion

It's remarkable that the individual log-scores in expressions (13) and (14) above are computationally expensive, but their sum results in the log-likelihood, which is less expensive.

The relation (13) invites us to 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 data and hypothesis locally, globally, and on all intermediate scales. To me this property makes the log-likelihood preferable to any single log-score (besides the fact that the log-likelihood is directly obtained from the principles of the probability calculus), because our interest is usually in how the hypothesis H relates to single data points as well as to any collection of them. I hope to discuss this point, which also involves the distinction between simple and composite hypotheses[15], more in detail elsewhere[16].

By applying the identity (8) and generalizing the expansion (9) to different divisions of the data – leave-two-out, leave-three-out, and so on – we see that the relation (13) can be generalized to any k-fold cross-validation log-scores. Thus the log-likelihood is also equivalent to an average of *all conceivable* cross-validation log-scores for all subsets of data, though I haven't calculated the weights of such average.

Thanks

To Aki Vehtari for some references. To the staff of the NTNU library for their always prompt assistance. To Mari, Miri, Emma for continuous encouragement and affection, and to Buster Keaton and Saitama for filling life with awe and inspiration. To the developers and maintainers

[15] Bernardo & Smith 2000 § 6.1.4. [16] Porta Mana 2019.

of Open Science Framework, LaTeX, Emacs, AUCTeX, Python, Inkscape, Sci-Hub for making a free and impartial scientific exchange possible.

This work is financially supported by the Kavli Foundation and the Centre of Excellence scheme of the Research Council of Norway (Roudi group).

Bibliography

Bernardo, J.-M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M., eds. (1985): *Bayesian Statistics 2*. (Elsevier and Valencia University Press, Amsterdam and Valencia). https://www.uv.es/~bernardo/valenciam.html.

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

Chandramouli, S. H., Shiffrin, R. M., Vehtari, A., Simpson, D. P., Yao, Y., Gelman, A., Navarro, D. J., Gronau, Q. F., et al. (2019): *Commentary on Gronau and Wagenmakers. Limitations of "Limitations of Bayesian leave-one-out cross-validation for model selection". Between the devil and the deep blue sea: tensions between scientific judgement and statistical model selection. Rejoinder: more limitations of Bayesian leave-one-out cross-validation*. Comput. Brain Behav. **2**^{1}, 12–47. See Gronau, Wagenmakers (2019).

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

Gelman, A., Hwang, J., Vehtari, A. (2014): *Understanding predictive information criteria for Bayesian models*. Stat. Comput. **24**^{6}, 997–1016.

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

— (1969): *A subjective evaluation of Bode's law and an 'objective' test for approximate numerical rationality*. J. Am. Stat. Assoc. **64**^{325}, 23–49. Partly repr. in Good (1983) ch. 13.

— (1975): *Explicativity, corroboration, and the relative odds of hypotheses*. Synthese **30**$^{1–2}$, 39–73. Partly repr. in Good (1983) ch. 15.

— (1981): *Some logic and history of hypothesis testing*. In: *Philosophy in economics*. Ed. by J. C. Pitt (Reidel), 149–174. Repr. in Good (1983) ch. 14 pp. 129–148.

— (1983): *Good Thinking: The Foundations of Probability and Its Applications*. (University of Minnesota Press, Minneapolis, USA).

— (1985): *Weight of evidence: a brief survey*. In: Bernardo, DeGroot, Lindley, Smith (1985), 249–270. With discussion by H. Rubin, T. Seidenfeld, and reply.

Gronau, Q. F., Wagenmakers, E.-J. (2019): *Limitations of Bayesian leave-one-out cross-validation for model selection*. Comput. Brain Behav. **2**^{1}, 1–11. See also comments and rejoinder in Chandramouli, Shiffrin, Vehtari, Simpson, Yao, Gelman, Navarro, Gronau, et al. (2019).

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).

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*, 3rd 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.

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

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).

Porta Mana, P. G. L. (2019): *Probabilistic models: models of what?* In preparation.

Stone, M. (1977): *An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion*. J. Roy. Stat. Soc. B **39**^1, 44–47.

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.