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

A relation between log-likelihood and cross-validation log-scores

(with some remarks on both)

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

Draft of 14 August 2019 (first drafted 8 August 2019)

***

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

1 Log-likelihoods

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

$$\mathrm{P}(H_h \mid D\, I) = \frac{\mathrm{P}(D \mid H_h\, I)\, \mathrm{P}(H_h \mid I)}{\mathrm{P}(D \mid I)} \tag{1a}$$

$$= \frac{\mathrm{P}(D \mid H_h\, I)\, \mathrm{P}(H_h \mid I)}{\sum_{h'} \mathrm{P}(D \mid H_{h'}\, I)\, \mathrm{P}(H_{h'} \mid I)}. \tag{1b}$$

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 us to assign the particular degrees of belief $\mathrm{P}(\dots \mid H\, I)$. In the latter case H is often called a '(probabilistic) model'.

The probability calculus, just like the truth calculus, proceeds purely syntactically rather than semantically. That is, if I tell you that H and $H \Rightarrow D$ are true, you can conclude that D is true; similarly, if I tell you that $\mathrm{P}(H \mid I) = p$ and $\mathrm{P}(H \Rightarrow D \mid I) = q$, you can conclude (try it as an exercise) that $\mathrm{P}(H\ D \mid I) = p + q - 1$. And in either case you don't need to know what H and D are about – they could be about Donald Duck or parallel universes. Don't we too often abuse of this syntactical property? We often say 'under model H our belief about the value of quantity x is expressed by such and such distribution $\mathrm{p}(x \mid H) = f(x)$', without explaining what θ really is and why it leads to f. Aren't terms such as 'model' and 'hypothesis', as often used in probability and statistics, convenient and respectable-looking carpets under which we can sweep the fact that we don't quite know what we're speaking about? The need to look under the carpet arises, though, the moment we have to specify our pre-data belief, the prior, about the mysterious H.

And yet again, semantics can very well be a by-product of syntax, or the distinction between the two be a chimera (Wittgenstein 1999; Girard 2001; 2003). Such important matters are unfortunately rarely discussed in probability and statistics.

Expression (1b) 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, the right side is only valid if

$$\mathrm{P}\big(\textstyle\bigvee_h H_h \mid I\big) = 1, \qquad \mathrm{P}(H_h \land H_{h'} \mid I) = 0 \quad \text{if } h \neq h'. \tag{2}$$

Only in extremely rare cases does the set of hypotheses $\{H_h\}$ encompass and reflect the extremely complex and fuzzy hypotheses lying in the backs of our minds. The background knowledge I is therefore only a simplified picture of our actual knowledge. That's why I or the hypotheses $\{H_h\}$ are often called *models*. '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' (Truesdell et al. 1980 Prologue p. xvi).

Expression (1a) 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 (1b). 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 (2003 §§ 4.3–4.4) insightful remarks about such binary comparisons, and also Good's (1950 § 6.3–6.6).

If our problem is to finally choose a hypothesis, discarding its competitors for future calculations, or more generally to make a decision (for example, choice of medical treatment) based on the observed data, the post-data belief (1) is necessary but not sufficient. We also need to specify a utility or cost function to calculate the expected gains of choosing one or another hypothesis or making one or another decision (Kadane et al. 1980; DeGroot 2004; Bernardo et al. 2000 ch. 2).

If our problem has an exploratory nature instead – for example, evaluating which hypotheses to include in our simplified set, or examining whether a hypothesis leads to peculiar beliefs for peculiar kinds of data – then all terms appearing in expression (1) are usually freely examined. In particular the term $\mathrm{P}(D \mid H_h\, I)$, called the *likelihood* of the hypothesis given the data (Good 1950 § 6.1 p. 62), or its logarithm

$$\log \mathrm{P}(D \mid H_h\, I), \tag{3}$$

in an arbitrary basis (Turing, Good (e.g. 1985; 1950; 1969), Jaynes (2003 § 4.2) 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* (Good 1950 ch. 6; 1975; 1981; 1985; and many other works in Good 1983; Osteyee et al. 1974 § 1.4; MacKay 1992; Kass et al. 1995; see also Jeffreys 1983 chs V, VI, A), are often used to get an idea of 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' (Osteyee et al. 1974 § 1.4.2 p. 12).

Recent literature (for example Kass et al. 1995) seems to exclusively deal with *relative* Bayes factors, so I'd like to mention that the non-relative Bayes factor for a hypothesis H_h provided by data D is actually defined as (Good 1981 § 2)

$$\frac{\mathrm{P}(D \mid H_h\, I)}{\mathrm{P}(D \mid \neg H_h\, I)} \equiv \frac{\mathrm{O}(H_h \mid D\, I)}{\mathrm{O}(H_h \mid I)} = \frac{\mathrm{P}(D \mid H_h\, I)\,[1 - \mathrm{P}(H_h \mid I)]}{\sum_{h'}^{h' \neq h} \mathrm{P}(D \mid H_{h'}\, I)\, \mathrm{P}(H_{h'} \mid I)}, \tag{4}$$

where the *odds* O is defined as $\mathrm{O} := \mathrm{P}/(1 - \mathrm{P})$. 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 don't enjoy. For example, the expected weight of evidence for a correct hypothesis is always positive, and for a wrong hypotheses always negative (Good 1950 § 6.7). See Jaynes (2003 §§ 4.3–4.4) for further discussion and a numeric example.

2 Cross-validation log-scores

The literature in probability and statistics has also employed various ad-hoc measures to make exploratory analyses. Here I consider one in particular: the *leave-one-out cross-validation log-score* (for example Stone 1977; Geisser et al. 1979; Bernardo et al. 1994 §§ 3.4, 6.1.6; Vehtari et al. 2002; 2012; Krnjajić et al. 2011; 2014; Gelman et al. 2014; Piironen et al. 2017; Gronau et al. 2019; Chandramouli et al. 2019), which I'll just call 'log-score' for brevity:

$$\frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid D_{-i}\, H_h\, I) \tag{5}$$

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 score, cursorily speaking, is 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.

Other variants of this score use more general partitions of the data into two disjoint subsets.

3 A relation

I'd like to show an exact relation between the log-score (5) and the log-likelihood (3) which doesn't seem to appear in the literature. I find this relation very intriguing: it says that *the log-likelihood is the sum of all averaged log-scores that can be formed from all data subsets*.

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

$$\mathrm{P}(D \mid H\, I) \equiv \big[\underbrace{\mathrm{P}(D \mid H\, I) \times \dots \times \mathrm{P}(D \mid H\, I)}_{d \text{ times}}\big]^{1/d} \tag{6}$$

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

$$\mathrm{P}(D \mid H\, I) = \mathrm{P}(D_i \mid D_{-i}\, H_h\, I) \times \mathrm{P}(D_{-i} \mid H_h\, I) \tag{7}$$

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 (6) using the product rule (7), using a different i for each of them. The result can be thus displayed:

$$
\begin{aligned}
\mathrm{P}(D \mid H\, I) \equiv \bigl[\, & \mathrm{P}(D_1 \mid D_{-1}\, H\, I) \times \mathrm{P}(D_{-1} \mid H\, I) \times \\
& \mathrm{P}(D_2 \mid D_{-2}\, H\, I) \times \mathrm{P}(D_{-2} \mid H\, I) \times \\
& \qquad\qquad \cdots \qquad\qquad\qquad \times \\
& \underset{\substack{\uparrow \\ \text{this column leads to the log-score}}}{\mathrm{P}(D_d \mid D_{-d}\, H\, I)} \times \mathrm{P}(D_{-d} \mid H\, I) \,\bigr]^{1/d} .
\end{aligned}
\tag{8}
$$

Upon taking the logarithm of this expression, the d factors vertically aligned on the left add up to the log-score (5), as indicated. But the mathematical reshaping we just did for $\mathrm{P}(D \mid H\, I)$ – that is, the root-product identity (6) and the expansion (8) – can be done for each of the remaining factors $\mathrm{P}(D_{-i} \mid H\, I)$ vertically aligned on the right in expression (8); and so on recursively. Here is an explicit example for $d = 3$:

$$
\begin{aligned}
&\mathrm{P}(D \mid H\, I) \equiv \\
&\quad \Bigl\{ \mathrm{P}(D_1 \mid D_2\, D_3\, H\, I) \times \bigl[\mathrm{P}(D_2 \mid D_3\, H\, I) \times \mathrm{P}(D_3 \mid H\, I) \times \\
&\qquad\qquad\qquad\qquad\qquad\quad \mathrm{P}(D_3 \mid D_2\, H\, I) \times \mathrm{P}(D_2 \mid H\, I) \bigr]^{1/2} \times \\
&\quad\;\; \mathrm{P}(D_2 \mid D_1\, D_3\, H\, I) \times \bigl[\mathrm{P}(D_1 \mid D_3\, H\, I) \times \mathrm{P}(D_3 \mid H\, I) \times \\
&\qquad\qquad\qquad\qquad\qquad\quad \mathrm{P}(D_3 \mid D_1\, H\, I) \times \mathrm{P}(D_1 \mid H\, I) \bigr]^{1/2} \times \\
&\quad\;\; \mathrm{P}(D_3 \mid D_1\, D_2\, H\, I) \times \bigl[\mathrm{P}(D_1 \mid D_2\, H\, I) \times \mathrm{P}(D_2 \mid H\, I) \times \\
&\qquad\qquad\qquad\qquad\qquad\quad \mathrm{P}(D_2 \mid D_1\, H\, I) \times \mathrm{P}(D_1 \mid H\, I) \bigr]^{1/2} \Bigr\}^{1/3} .
\end{aligned}
\tag{9}
$$

In this example, the logarithm of the three vertically aligned factors in the left column is, as already noted, the log-score (5). 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

$$\begin{aligned}
\log \mathrm{P}(D \mid H\, I) \equiv{}& \frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid D_{-i}\, H\, I) + \\
& \frac{1}{d}\sum_{i\in\{1,\dots,d\}} \frac{1}{d-1} \sum_{j\in\{1,\dots,d\}}^{j\neq i} \log \mathrm{P}(D_{-i,j} \mid D_{-i,-j}\, H\, I) + \\
& \binom{d}{d-2}^{-1} \sum_{i,j\in\{1,\dots,d\}}^{i<j} \frac{1}{d-2} \sum_{k\in\{1,\dots,d\}}^{k\neq i,j} \log \mathrm{P}(D_{-i,-j,k} \mid D_{-i,-j,-k}\, H\, I) + \\
& \cdots + \\
& \binom{d}{2}^{-1} \sum_{i,j\in\{1,\dots,d\}}^{i<j} \frac{1}{2}\bigl[\log \mathrm{P}(D_i \mid D_j\, H\, I) + \log \mathrm{P}(D_j \mid D_i\, H\, I)\bigr] + \\
& \frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid H\, I),
\end{aligned} \tag{10}$$

which can be compactly written

$$\boxed{\log \mathrm{P}(D \mid H\, I) \equiv \sum_{k=1}^{d} \binom{d}{k}^{-1} \sum_{\substack{\text{ordered}\\ k\text{-tuples}}} \frac{1}{k} \sum_{\substack{\text{cyclic}\\ \text{permutations}}} \log \mathrm{P}(D_{i_1} \mid D_{i_2} \cdots D_{i_k}\, H\, I).} \tag{11}$$

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 the kth-order log-scores being 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

$$\boxed{\log \mathrm{P}(D \mid H\, I) \equiv \frac{1}{d} \sum_{j=1}^{d} \sum_{k=0}^{d-1} \binom{d-1}{k}^{-1} \sum_{\substack{\text{ordered}\\ k\text{-tuples,}\\ j \text{ excluded}}} \log \mathrm{P}(D_j \mid D_{i_1} \cdots D_{i_k}\, H\, I).} \tag{12}$$

4 Discussion

The relation (11) just proven between the log-likelihood and the log-score brings forth several thoughts.

It's remarkable that the individual log-scores in expressions (11) and (12) above are computationally expensive, but their sum results in a less expensive quantity: the log-likelihood.

The relation (11) 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 hypothesis and data locally, globally, and on all intermediate scales.

To me this makes sense, because our interest is in how the hypothesis relates to single data points as well as to groups of them. A good portion of recent literature seems to focus on the degrees of belief about data given some hypothesis only *after* some data have been collected (O'Hagan 1995; Berger et al. 1996; 1998).

If we see the symbol H as just a placeholder for a probabilistic model, rather than an empirical hypothesis (see the remark about syntax and semantics on p. 1), we must remember that the conjunction of H and some data D simply defines a new probabilistic model $H' \coloneqq H\, D$.

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

$$\mathrm{P}(D \mid D'\, \hat{H}\, I) = \mathrm{P}(D \mid \hat{H}\, I) \qquad \text{if } D' \nRightarrow D, \tag{13}$$

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

$$\mathrm{P}(D D' \mid H\, I) = \sum_h \mathrm{P}(D \mid \hat{H}_h\, H\, I)\, \mathrm{P}(D' \mid \hat{H}_h\, H\, I)\, \mathrm{P}(\hat{H}_h \mid H\, I) \qquad \text{if } D' \nRightarrow D. \tag{14}$$

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

$$\frac{1}{d}\sum_{i=1}^{d} \log \mathrm{P}(D_i \mid D_{-i}\, H\, I) \approx \frac{1}{d} \log \mathrm{P}(D \mid H\, I). \tag{15}$$

This is in fact an exact equality if property (13) 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)

*** remark that $D\, H$ is a *new* probability model: which of the two are we assessing? Connection with learning vs non-learning models which I hope to take up in another work.

*** (Bernardo et al. 1994 § 6.1.6) show the approximation only valid if number of data large enough – not very interesting situation in today's problems.

*** 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.)

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., 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. (1994): *Bayesian Theory*. (Wiley, Chichester).

— (2000): *Bayesian Theory*, reprint. (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).

Curien, P.-L. (2001): *Preface to Locus solum*. Math. Struct. in Comp. Science **11**3, 299–300. See also Girard (2001).

DeGroot, M. H. (2004): *optimal statistical decisions*, reprint. (Wiley, New York).

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.

Girard, J.-Y. (2001): *Locus solum: From the rules of logic to the logic of rules*. Math. Struct. in Comp. Science **11**3, 301–506. http://iml.univ-mrs.fr/~girard/Articles.html. See also Curien (2001).

— (2003): *From foundations to ludics*. Bull. Symbolic Logic **9**2, 131–168. http://iml.univ-mrs.fr/~girard/Articles.html.

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.

Good, I. J. (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).

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.

Kadane, J. B., Dickey, J. M. (1980): *Bayesian decision theory and the simplification of models*. In: *Evaluation of econometric models*. Ed. by J. Kmenta, J. B. Ramsey (Academic Press, New York), 245–268. https://www.nber.org/chapters/c11704.

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.

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

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

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

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

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.

Truesdell III, C. A., Muncaster, R. G. (1980): *Fundamentals of Maxwell's Kinetic Theory of a Simple Monatomic Gas: Treated as a Branch of Rational Mechanics*. (Academic Press, New York).

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.

Wittgenstein, L. (1999): *Philosophical Investigations / Philosophische Untersuchungen*, reissued second ed. (Blackwell, Oxford). German text, with English transl. by G. E. M. Anscombe. Written 1945–49; first publ. 1953.