# The relation between leave-one-out log-scores and log-evidence

P.G.L. Porta Mana
<piero.mana@ntnu.no>

Draft of 8 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 tells us unequivocally how our degree of belief in a hypothesis $H_h$ given data $D$ and knowledge $K$ is related to our degree of belief in observing those data when we entertain that hypothesis as true:

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

The hypothesis is one in a set of mutually exclusive and exhaustive hypotheses $\{H_h\}$. Let's call $\mathrm{P}(H_h \mid D\, K)$ the post-data probability or belief in $H_h$. If our pre-data beliefs $\mathrm{P}(H_h \mid K)$ in the hypotheses all equal, then the ratio of the post-data probabilities for two hypotheses is equal to the ratio of the beliefs in the data given the hypotheses:

$$\frac{\mathrm{P}(H_h \mid D\, K)}{\mathrm{P}(H_{h'} \mid D\, K)} = \frac{\mathrm{P}(D \mid H_h\, K)}{\mathrm{P}(D \mid H_{h'}\, K)} \qquad \text{(if } \mathrm{P}(H_h \mid K) = \mathrm{P}(H_{h'} \mid K)\text{)}. \tag{2}$$

Such ratio or its logarithm is variously called *weight of evidence* and *Bayes factor* (Good 1950; Osteyee et al. 1974; 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).

The literature in probability and statistics has also employed, for quite some time, 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* (Krnjajić et al. 2011; Geisser et al. 1979; Vehtari et al. 2002; 2012; Draper et al. 2014; 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 $D_d$ is a datum in the data $D$, and $D_{-d}$ represents the data with datum $D_d$ excluded. The intuition behind this measure is more or less this: 'let's see what my belief in the last datum should be, on average, once I've observed the other data, if I consider $H_h$ as true'. On average means considering all the different orders in which the data could have been observed.

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

My point of view, however, 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 log-score above can be viewed as an approximation of the logarithm of the post-data probability (1) is actually a refined ***

$$\begin{aligned}
\bigl[\mathrm{P}(D_1 \mid D_{-1}\, H_h\, K) \times \mathrm{P}(D_{-1} \mid H_h\, K) \times{}& \\
\mathrm{P}(D_2 \mid D_{-2}\, H_h\, K) \times \mathrm{P}(D_{-2} \mid H_h\, K) \times{}& \\
\dots \qquad\qquad \times{}& \\
\mathrm{P}(D_d \mid D_{-d}\, H_h\, K) \times \mathrm{P}(D_{-} \mid H_h\, K)\bigr]^{\frac{1}{d}}&
\end{aligned} \tag{4}$$

# Bibliography

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

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).

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).

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

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.