International Journal of Statistics and Probability; Vol. 7, No. 1; January 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Characterization of the Bayesian Posterior Distribution in Terms of Self-information

Marco Dall'Aglio[1] & Theodore P. Hill[2]

[1] Department of Economics and Finance, Luiss University, Italy

[2] School of Mathematics, Georgia Institute of Technology & California Polytechnic State University, USA

Correspondence: Marco Dall'Aglio, Department of Economics and Finance, Luiss University, Rome, Italy.
E-mail: mdallaglio@luiss.it

Received: October 5, 2017 Accepted: October 20, 2017 Online Published: November 6, 2017

doi:10.5539/ijsp.v7n1p21 URL: https://doi.org/10.5539/ijsp.v7n1p21

Abstract

It is well known that the classical Bayesian posterior arises naturally as the unique solution of different optimization problems, without the necessity of interpreting data as conditional probabilities and then using Bayes' Theorem. Here it is shown that the Bayesian posterior is also the unique minimax optimizer of the loss of self-information in combining the prior and the likelihood distributions, and is the unique proportional consolidation of the same distributions. These results, direct corollaries of recent results about conflations of probability distributions, further reinforce the use of Bayesian posteriors, and may help partially reconcile some of the differences between classical and Bayesian statistics.

Keywords: Bayesian analysis, conflation of probability distributions, likelihood ratio, Shannon information.

1. Introduction

In statistics, prior belief about the value of an unknown parameter, $\theta \in \Theta \subseteq \mathbb{R}^n$ obtained from experiments or other methods, is often expressed as a Borel probability distribution $P_0(\cdot)$ on $\Theta \subseteq \mathbb{R}^n$ called the *prior distribution*. New evidence or information about the value of θ, based on an independent experiment or survey, i.e. a random variable X, is recorded as a *likelihood* $L(\cdot) = p(\cdot|X)$, the conditional distribution of θ given an observable X. Given the prior distribution P_0 and the likelihood distribution L, a *posterior distribution* $P_1 = P_1(P_0, L)$ for θ incorporates the new likelihood information about θ into the information from the prior, thereby updating the prior.

Bayesian and likelihood inference in general does not require the prior and/or the likelihood to be normalizable. This is the case, for instance, of improper priors, that are often used to convey the notion of lack of prior information about the parameter. Similarly, the likelihood, while conceived as a parametric family of probability distributions over the data, does not even require, in principle, to be measurable w.r.t. the parameter space.

We will assume the prior P_0 and the likelihood L to be non-negative measures[1]. Such measures will be discrete, yielding a mass function[2] (p.m.f.), or absolutely continuous (w.r.t. the Lebesgue measure), yielding a density function (p.d.f.). In both cases we will require the prior and the likelihood to be *compatible*, i.e. such that the measure defined by the products of the p.m.f.'s (p.d.f.'s, respectively) is normalizable. Here and throughout we will assume θ to be real-valued (with generalizations to the multidimensional framework left to the interested reader).

In the classical framework, the posterior distribution P_1 is the Bayes posterior distribution obtained as the conditional distribution of θ given the new likelihood information, but the same Bayes posterior distribution has also been derived in several information-theoretic contexts. Shore and Johnson (1980) give axiomatic foundations for deriving various probabilistic rules and, more specifically, the combining mechanism for the Bayes rule in Bernardo (1979) is expected utility, in Zellner (1988) is an information processing rule, and in Zellner (1996) is a maximum entropy principle. More recently, the self information loss, together with the Kullback-Leibler divergence, has been employed in a proper Bayesian setting to derive objective prior distributions for specific discrete parameter spaces (Villa and Walker, 2015) and to estimate the number of degrees of freedom in the t-distribution (Villa and Walker, 2014).

The main goal of this note is to complement those characterizations by applying recent results for conflations of probability distributions (Hill, 2011) to show that the Bayesian posterior is the unique posterior that minimizes the maximum loss of self-information in combining the prior and likelihood distributions. Secondary goals are to show that the Bayesian

[1] L is usually called a *likelihood function* to mark the fact that it may be non-normalizable. In this context where improper distributions are included, we will refer to L as a likelihood distribution (proper or improper.)

[2] For simplicity, we will adopt the acronym p.m.f. (probability mass function) to indicate the measure of the single atoms in a discrete distributions even when the distribution is improper. Similarly, we will indicate the density of an a.c. distribution, proper or improper, with the acronym p.d.f.

posterior is the unique posterior that is a proportional consolidation of the prior and likelihood distributions. Another direct corollary of recent results for conflations of probability distributions (Hill and Miller, 2011), the problem of identifying the best posterior when the prior and likelihood distributions are not weighted equally is addressed, complementing results in Zellner (2002). This new weighted posterior, the unique distribution that minimizes the maximum loss of weighted self-information, coincides with the classical Bayesian posterior if the prior and likelihood are weighted equally, but in general is different. We conclude with an open question regarding the minimax likelihood ratio of the prior and likelihood distributions.

2. Combining Priors and Likelihoods into Posteriors

There are many different methods for combining several probability distributions (e.g., see Genest and Zidek (1986); Hill (2011)), and in particular, for combining the prior distribution P_0 and the likelihood distribution L into a single *posterior distribution* $P_1 = P_1(P_0, L)$. For example, the prior and likelihoods could simply be averaged, i.e. $P_1 = \frac{P_0+L}{2}$, perhaps reflecting additional knowledge that the prior and likelihood distributions resulted from two different independent experiments, only one of which is assumed to be the "correct" experiment, and it is not known which.

The classical *Bayesian posterior distribution* P_B is defined via Bayes Theorem: if P_0 and L are discrete with p.m.f.'s p_0 and p_L respectively, then P_B is discrete with p.m.f.

$$p_B(\theta) = \frac{p_0(\theta)p_L(\theta)}{\sum_{\widehat{\theta}\in\Theta} p_0(\widehat{\theta})p_L(\widehat{\theta})};$$

and if P_0 and L are absolutely continuous with probability density functions (p.d.f.'s) f_0 and f_L respectively, then P_B is absolutely continuous with p.d.f.

$$f_B(\theta) = \frac{f_0(\theta)f_L(\theta)}{\int_\Theta f_0(\widehat{\theta})f_L(\widehat{\theta})d\widehat{\theta}}.$$

The same results hold true for improper prior or likelihood distributions, provided the denominators are positive and finite.

3. Minimax Loss of Self-information

When the goal is to consolidate information from a prior distribution and a likelihood distribution into a (posterior) distribution, replacing those two distributions by a single distribution will clearly result in some loss of information, however that is defined. Recall that the *self-information* (also called the *surprisal* or *Shannon information*, Shannon (1948)) of the random event A, $S_P(A)$, is given by $S_P(A) = -\log_2 P(A)$. (N.B. The *Shannon entropy* of a probability, on the other hand, is the *expected value* of the self-information, and in some contexts the terms surprisal or self-information are also used to mean this expected value entropy context.) The numerical value of the self-information of a given event is simply the number of binary bits of information reflected in its probability (so the smaller the value of $P(A)$, the greater the information or surprise).

Example 3.1. If P is uniformly distributed on $(0, 1)$ and $A = (0, 0.25) \cup (0.5, 0.75)$, then the self-information of A is $S_P(A) = -\log_2(P(A)) = -\log_2(0.5) = 1$, so if X is a random variable with distribution P, then exactly one binary bit of information is obtained by observing that $X \in A$, in this case that the value of the second binary digit of X is 0.

Definition 3.2. The *combined self-information* associated with the event A under the prior distribution P_0 and the likelihood distribution L is

$$S_{\{P_0,L\}}(A) = -\log_2 P_0(A)L(A).$$

Note that when $P(A)$ is finite, the combined self-information is simply the sum of the self-informations under the prior and likelihood distributions, and that this is the self-information of the event that A is observed independently under both the prior and the likelihood distributions.

Similarly, the *maximum loss between the self-information of a posterior distribution P_1 and the combined self-information* of the prior and likelihood distributions P_0 and L, $M(P_1; P_0, L)$, is

$$M(P_1; P_0, L) = \max_A \left\{\log_2 \frac{P_1(A)}{P_0(A)L(A)}\right\}.$$

In the case of improper distributions we will assume that, when $P_1(A) = \infty$, and either $P_0(A) = \infty$ or $L(A) = \infty$ (or both), the ratio is 1. Instead, when all distributions are proper, the quantity to be maximized in A is the difference between the combined self-information associated with the event A under the prior P_0 and the likelihood L, and the self-information of P_1 associated with the same event.

Definition 3.3. A prior distribution P_0 and a likelihood distribution L, both proper or improper, are *compatible* if P_0 and L are both discrete with p.m.f.'s p_0 and p_L satisfying $0 < \sum_{\theta\in\Theta} p_0(\theta)p_L(\theta) < \infty$, or are both absolutely continuous with p.d.f.'s f_0 and f_L satisfying $0 < \int_\Theta f_0(\theta)f_L(\theta)d\theta < \infty$.

Example 3.4. Every two geometric distributions are compatible, every two normal distributions are compatible, and every exponential distribution is compatible with every normal distribution. Also, when improper priors are considered, they are chosen to be compatible with the likelihood. Distributions with disjoint support, discrete or continuous, are not compatible.

Remark. In practice, compatibility is not problematic when both P_0 and L are proper. Any two distributions may be easily transformed into two new distributions, arbitrarily close to the original distributions, so that the two new distributions are compatible, for instance by convolving each with a $N(0, \epsilon)$ distribution.

Theorem 3.5. *Let P_0 and L be proper or improper discrete compatible prior and likelihood distributions. Then the Bayesian posterior P_B is the unique proper or improper posterior distribution that minimizes the maximum loss of self-information from the prior and likelihood distributions, i.e., that minimizes $M(P_1; P_0, L)$ among all posterior distributions P_1. Moreover,*

$$M(P_1; P_0, L) \geq \log_2 \left[\left(\sum_{\theta\in\Theta} p_0(\theta)p_L(\theta) \right)^{-1} \right] \text{ for all posterior distributions } P_1,$$

and equality is uniquely attained by the Bayesian posterior $P_1 = P_B$.

Proof. Since $\log_2(x)$ is strictly increasing, the maximum loss between the self-information of a posterior distribution P_1 and the combined self-information of the prior and likelihood distributions P_0 and L, occurs for an event A where $\frac{P_1(A)}{P_0(A)L(A)}$ is maximized. Clearly, $M_1(P_1; P_0, L) = \infty$ whenever P_1 is improper, while P_0 and L are compatible. Since our goal is to minimize $M(P_1; P_0, L)$, we restrict our search for the optimal P_1 to proper distributions. The conclusion of Theorem 3.5 then follows as an application of Hill (2011, Corollary 4.4) where it is shown that the conflation of two discrete Borel probability distribution is the unique Borel probability distribution that minimizes the maximum loss of Shannon information between those distributions. It turns out that the Bayesian posterior is the conflation of the prior and the likelihood distributions. Consequently, the lower bound for the maxmin loss of information, valid for the conflation of any finite number of discrete Borel probability distributions, can be applied to the Bayesian paradigm as well. Analogous results hold (see Hill (2011, Theorem 4.5)) for a.c. distributions. □

4. Proportional Posteriors

Another criterion to assess the quality of the posterior distribution is to require that it reflects the relative likelihoods of identical individual outcomes under both P_0 and L. For example, if the probability that the prior and the (independent) likelihood are both θ_a is twice that of the probability both are θ_b, then $P_1(\theta_a)$ should also be twice as large as $P_1(\theta_b)$.

Definition 4.1. A discrete (posterior) distribution P^*, proper or improper, with p.m.f. p^* is a *proportional posterior of a discrete prior distribution P_0 with p.m.f. p_0 and a compatible discrete likelihood distribution L with p.m.f. p_L*, both proper or improper, if

$$\frac{p^*(\theta_a)}{p^*(\theta_b)} = \frac{p_0(\theta_a)p_L(\theta_a)}{p_0(\theta_b)p_L(\theta_b)} \quad \text{for all } \theta_a, \theta_b \in \Theta.$$

Similarly, a proper or improper posterior a.c. distribution P^* with p.d.f. f^* is a *proportional posterior of an a.c. prior distribution P_0 with p.d.f. f_0 and a compatible likelihood distribution L with p.d.f. f_L*, both proper or improper, if

$$\frac{f^*(\theta_a)}{f^*(\theta_b)} = \frac{f_0(\theta_a)f_L(\theta_a)}{f_0(\theta_b)f_L(\theta_b)} \quad \text{for (Lebesgue) almost all } \theta_a, \theta_b \in \Theta.$$

Theorem 4.2. *Let P_0 and L be two proper or improper, compatible discrete or compatible absolutely continuous prior and likelihood distributions, respectively. Then the Bayesian posterior distribution P_B is a proportional consolidation for P_0 and L.*

Proof. A result from (2011, Theorem 5.5) shows that the conflation of two probability distributions is the unique proper proportional consolidation of those distributions. Consequently, the Bayesian posterior is the unique proper proportional consolidation of P_0 and L. No improper distribution shares the same property. In fact, if all the distributions are discrete, and Q is an improper proportional consolidation of P_0 and L with p.m.f. q, then $q(\theta_1) = kp_0(\theta_1)L(\theta_1)$ for some $\theta_1 \in \Theta$ and $k > 0$, with $k \neq 1$. Since Q is a proportional consolidation, then

$$q(\theta) = \frac{p_0(\theta)p_L(\theta)}{p_0(\theta_1)p_L(\theta_1)} q(\theta_1) = kp_0(\theta)p_L(\theta) \quad \text{for every } \theta \in \Theta$$

Summing over all Θ we obtain a finite mass for Q – a contradiction. A similar proof works for a.c. distributions. □

5. Optimal Posteriors for Weighted Prior and Likelihood Distributions

Definition 5.1. Given a prior distribution P_0 with weight $w_0 > 0$ and a likelihood distribution L with weight $w_L > 0$, the *combined weighted self-information* associated with the event A, $S_{(P_0,w_0;L,w_L)}(A)$, is

$$S_{(P_0,w_0;L,w_L)}(A) = \frac{w_0}{\max\{w_0, w_L\}} S_{P_0}(A) + \frac{w_L}{\max\{w_0, w_L\}} S_L(A).$$

This definition ensures that only the *relative weights* are important, so for instance if $w_0 = w_L$, the combined weighted self-information of the prior and likelihood always coincides with the (unweighted) combined self-information of the prior and likelihood. The next theorem, a special case of Hill and Miller (2011, (8)), identifies the posterior distribution that minimizes the loss of weighted self-information in the case the prior and likelihood distributions are compatible discrete distributions; the case for compatible absolutely continuous distributions is analogous.

Theorem 5.2. *Let P_0 and L be compatible discrete prior and likelihood distributions, proper or improper, with p.m.f.'s p_0 and p_L and weights $w_0 > 0$ and $w_L > 0$, respectively. Then the unique posterior distribution that minimizes the maximum loss of self-information from the weighted prior and likelihood distributions, i.e., that minimizes, among all posterior distributions P_1, proper or improper, the difference between the combined weighted self-information of the prior and the likelihood distributions and the self-information of the posterior, i.e.*

$$\max_A \left\{ \frac{P_1(A)}{P_0(A)^{\frac{w_0}{\max\{w_0,w_L\}}} L(A)^{\frac{w_L}{\max\{w_0,w_L\}}}} \right\},$$

is the posterior distribution P_1^w with p.m.f.

$$p_1^w(\theta) = \frac{(p_0(\theta))^{\frac{w_0}{\max\{w_0,w_L\}}} (p_L(\theta))^{\frac{w_L}{\max\{w_0,w_L\}}}}{\sum_{\hat{\theta}\in\Theta} (p_0(\widehat{\theta}))^{\frac{w_0}{\max\{w_0,w_L\}}} (p_L(\widehat{\theta}))^{\frac{w_L}{\max\{w_0,w_L\}}}}.$$

Remark. If both the prior and likelihood distributions are normally distributed, the Bayesian posterior is also a best linear unbiased estimator (BLUE) and a maximum likelihood estimator (MLE); e.g. see Hill (2011).

6. An Open Question

In classical hypotheses testing, a standard technique to decide from which of several known distributions given data actually came is to maximize the likelihood ratios, that is, the ratios of the p.m.f.'s or p.d.f.'s. Analogously, when the objective is to decide how best to consolidate a prior distribution P_0 and a likelihood distribution L into a single (posterior) distribution $P_1 = P_1(P_0, L)$, one natural criterion is to choose P_1 so as to make the ratios of the likelihood of observing θ under P_1 as close as possible to the likelihood of observing θ under *both* the prior distribution P_0 and the likelihood distribution L. This motivates the following notion of minimax likelihood ratio posterior.

Definition 6.1. A proper discrete probability distribution P^* (with p.m.f. p^*) is the *minimax likelihood ratio (MLR) posterior* of a discrete prior distribution P_0 with p.m.f. p_0 and a discrete likelihood distribution L with p.m.f. p_L, compatible with each other and both proper or improper, if

$$\min_{\text{p.m.f.'s } p} \left\{ \max_{\theta\in\Theta} \frac{p(\theta)}{p_0(\theta)p_L(\theta)} - \min_{\theta\in\Theta} \frac{p(\theta)}{p_0(\theta)p_L(\theta)} \right\}$$

is attained by $p = p^*$ (where $0/0 := 1$).

Similarly, a proper a.c. distribution P^* with p.d.f. f^* is the MLR posterior of an a.c. prior distribution P_0 with p.d.f. f_0 and an a.c. likelihood distribution L with p.d.f. f_L, compatible with each other and both proper or improper, if

$$\min_{\text{p.m.f.'s } f} \left\{ \operatorname*{ess\,sup}_{\theta\in\Theta} \frac{f(\theta)}{f_0(\theta)f_L(\theta)} - \operatorname*{ess\,inf}_{\theta\in\Theta} \frac{f(\theta)}{f_0(\theta)f_L(\theta)} \right\}$$

is attained by f^*.

The min-max terms in Definition 6.1 are similar to the min-max criterion for loss of self-information (Theorem 3.5), whereas the others are dual max-min criteria. Hill (2011, Theorem 5.2), can be used to prove that, when P_0 and L are both proper, the Bayesian posterior is the unique MLR consolidation of the prior and likelihood distributions among all proper Borel distributions. Whether the same result can be extended to prove that the Bayesian posterior is the unique MLR consolidation among both proper and improper distributions remains an open question.

Acknowledgement

The second author is grateful to LUISS University in Rome for its generous invitation to visit for a month, for its warm hospitality during that stay, and for the lively discussions there that were the genesis of these ideas.

References

Bernardo, J. M. (1979a). Expected information as expected utility, *Ann. Statist.* **7**, 686–690.

Ebrahimi, N., Soofi, E. S., & Soyer, R. (2010). Information Measures in Perspective, *International Statistical Review* **78**(3), 383–412.

Genest, C., & Zidek, J. (1986). Combining probability distributions: a critique and an annotated bibliography, *Statist. Sci.* **1**, 114–148. https://doi.org/10.1214/ss/1177013825

Hill, T. (2011). Conflations of Probability Distributions, *Transactions of the American Mathematical Society* **363**(6), 3351–3372.

Hill, T., & Miller, J. (2011). How to combine independent data sets for the same quantity, *Chaos* **21**. http://doi.org/10.1063/1.3593373.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell Sys. Tech. J.* **27**, 379–423.

Shore, J. E., & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, *IEEE Trans. Inform. Theory* **26**, 26–37.

Villa, C., & Walker, S. (2014). Objective priors for the number of degrees of freedom of a *t*-distribution, *Bayesian Anal.* **9**(1), 197–220.

Villa, C., & Walker, S. (2015). An objective approach to prior mass functions for discrete parameter spaces, *J. Amer. Statist. Assoc* **120**(511), 1072–1082. https://doi.org/10.1080/01621459.2014.946319

Zellner, A. (1988). Optimal information processing and Bayes' Theorem, *Amer. Statist.* **42**, 278–284 (with discussion).

Zellner, A. (1996). Bayesian method of moments/instrumental variable (BMOM/IV) analysis of mean and regression models. In *Prediction and Modelling Honoring Seymour Geisser*, Eds. J. C. Lee, A. Zellner and W. O. Johnson, 61–74, Netherlands: Springer-Verlag.

Zellner, A. (2002). Information processing and Bayesian analysis, *J. Econometrics* **107**. https://doi.org/10.1016/S0304-4076(01)00112-9 41–50.