LUISS

IRIS - Luiss Guido Carli Research Institutional Repository | Catalogo istituzionale della ricerca Luiss Guido Carli

Bayesian Posteriors Without Bayes' Theorem

Post-print version of the following publication: | Versione post-print della seguente pubblicazione:

Original Citation/Citazione:
Dall'Aglio, Marco; Theodore P., Hill. (2012). Bayesian Posteriors Without Bayes' Theorem. no. 1203.0251. http://arxiv.org/abs/1203.0251

Availability/Disponibilità:
This version is available at: 11385/56256 since: 2015-11-09T10:17:33Z - Questa versione è disponibile alla pagina: 11385/56256 dal: 2015-11-09T10:17:33Z

Publisher/Casa editrice:

Published version/Pubblicato:

License/Licenza:
Attribution-NonCommercial-NoDerivatives 4.0 International

Availability/Termini d'uso:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. Works made available under a Creative Commons license can be used according to the terms and conditions of said license. For all terms of use and more information see the publisher's website. | I termini e le condizioni relativi al riutilizzo della presente versione della pubblicazione sono disciplinati dalla politica editoriale. Le opere messe a disposizione con licenze Creative Commons possono essere utilizzate conformemente ai termini e alle condizioni previste da tali licenze. Per l'insieme delle condizioni di utilizzo e per ulteriori informazioni si rinvia al sito web dell'editore.

This item was downloaded from IRIS Luiss (https://iris.luiss.it/). When citing, please refer to the published version. | Questo documento è stato scaricato da IRIS Luiss (https://iris.luiss.it/). Per la citazione, fare riferimento alla versione pubblicata sul sito dell'editore.

(Article begins on next page | Il contributo inizia nella pagina successiva)

11 October 2026

arXiv:1203.0251v1 [math.ST] 1 Mar 2012

Bayesian Posteriors Without Bayes' Theorem

T. P. Hill and Marco Dall' Aglio

Abstract

The classical Bayesian posterior arises naturally as the unique solution of several different optimization problems, without the necessity of interpreting data as conditional probabilities and then using Bayes' Theorem. For example, the classical Bayesian posterior is the unique posterior that minimizes the loss of Shannon information in combining the prior and the likelihood distributions. These results, direct corollaries of recent results about conflations of probability distributions, reinforce the use of Bayesian posteriors, and may help partially reconcile some of the differences between classical and Bayesian statistics.

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 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, is recorded as a *likelihood distribution* L. Here and throughout it will be assumed that the likelihood function has finite positive total mass, and that L has been normalized, so that in fact L is also a Borel probability distribution on Θ. 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, thus updating the prior. The posterior distribution P_1 is typically viewed as the conditional distribution of θ given the new likelihood information, often expressed as a random variable X.

The first main goal of this note is to use recent results for conflations of probability distributions [3, 4] to show that the Bayesian posterior is the unique posterior that minimizes the loss of Shannon information in combining the prior and likelihood distributions. The Bayesian posterior is also the unique posterior that attains the minimax likelihood ratio of the prior and likelihood distributions, and the unique posterior that is a proportional consolidation of the prior and likelihood distributions. Thus, the classical Bayesian posterior appears naturally as the solution of several different optimization problems, without the necessity of interpreting likelihood as a conditional probability and then invoking Bayes Theorem. These results reinforce the use of Bayesian posteriors, and may help partially reconcile some of the differences between classical statistics and Bayesian statistics.

The second main goal of this note, another direct corollary of recent results for conflations of probability distributions [4], is to identify the best posterior when the prior and likelihood distributions are not weighted equally, such as in cases when the prior distribution is

given more weight than the likelihood distribution. This new weighted posterior, the unique distribution that minimizes the loss of weighted Shannon information, coincides with the classical Bayesian posterior if the prior and likelihood are weighted equally, but in general is different.

2 Combining Priors and Likelihoods into Posteriors

There are many different methods for combining several probability distributions (e.g., see [1, 3]), 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. $P_1 = \frac{P_0+L}{2}$, or the data underlying the prior and the likelihood could be averaged, in which case the posterior P_1 would be the distribution of $\frac{X_0+X_L}{2}$, where X_0 and X_L are independent random variables with distributions P_0 and L, respectively.

In Bayesian statistics, the likelihood function L is usually interpreted as $L(\theta) = \alpha P(X \mid \theta)$, where X is the independent experiment or random variable yielding new information about θ, and α is the normalizing constant for L to have mass one (cf. [2]). The *Bayesian posterior distribution* P_B is then calculated using Bayes Theorem: for example, if P_0 and L are discrete with probability mass functions (p.m.f.'s) p_0 and p_L respectively, then P_B is discrete with p.m.f.

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

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.

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

(provided the denominators are positive and finite).

3 Minimizing Loss of Shannon 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 classical *Shannon information* (also called the *self-information* or *surprisal*) associated with the event A from a probability distribution P, $S_P(A)$, is given by $S_P(A) = -\log_2 P(A)$ (so the smaller the value of $P(A)$, the greater the information or surprise). The numerical value of the Shannon information of a given probability is simply the number of binary bits of information reflected in that probability.

Example 3.1. If P is uniformly distributed on $(0, 1)$ and $A = (0, 0.25) \cup (0.5, 0.75)$, then $S_P(A) = -\log_2(P(A)) = -\log_2(0.5) = 1$, 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 Shannon Information* associated with the event A from the prior distribution P_0 and the likelihood distribution L is

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

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

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

Note that the definition of combined Shannon information implicitly assumes independence of the prior and likelihood distributions. Note also that no information is obtained by observing an event that is certain to occur, so for instance $S_{[P_0,L]}(\Theta) = S_{P_1}(\Theta) = 0$. This implies that $M(P_1; P_0, L)$ is never negative.

Definition 3.3. A prior distribution P_0 and a likelihood distribution L are *compatible* if P_0 and L are both discrete with p.m.f's p_0 and p_L satisfying $\sum_{\theta\in\Theta} p_0(\theta)p_L(\theta) > 0$, 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. Distributions with disjoint support, discrete or continuous, are not compatible.

Remark. In practice, compatibility is not problematic. 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 $U(-\epsilon, \epsilon)$ distribution.

Theorem 3.5. *Let P_0 and L be discrete compatible prior and likelihood distributions. Then the Bayesian posterior P_B is the unique posterior distribution that minimizes the maximum loss of Shannon 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 $P_1 = P_B$.

The conclusion of Theorem 3.5 follows immediately as a special case of [3, Corollary 4.4]; analogous conclusions for the case of compatible absolutely continuous distributions follow from [3, Theorem 4.5]. For the benefit of the reader, a sketch of the proof of Theorem 3.5 similar to that in [4] is included.

Sketch of proof. First observe that for an event A, the difference between the combined Shannon information obtained from a prior distribution P_0 and a likelihood distribution L, and the Shannon information obtained from the posterior P_1, is

$$S_{\{P_0,L\}}(A) - S_{P_1}(A) = S_{P_0}(A) + S_L(A) - S_{P_1}(A) = \log_2 \frac{P_1(A)}{P_0(A)L(A)}.$$

Since $\log_2(x)$ is strictly increasing, the maximum (loss) thus occurs for an event A where $\frac{P_1(A)}{P_0(A)L(A)}$ is maximized.

Next note that the largest loss of Shannon information occurs for small sets A, since for disjoint sets A and B,

$$\frac{P_1(A \cup B)}{P_0(A \cup B)L(A \cup B)} \leq \frac{P_1(A) + P_1(B)}{P_0(A)L(A) + P_0(B)L(B)} \leq \max\left\{ \frac{P_1(A)}{P_0(A)L(A)}, \frac{P_1(B)}{P_0(B)L(B)} \right\},$$

where the inequalities follow from the inequalities $(a+b)(c+d) \geq ac + bd$ and $\frac{a+b}{c+d} \leq \max\left\{\frac{a}{c}, \frac{b}{d}\right\}$ for positive numbers a, b, c, d. Thus the problem reduces to finding the probability mass function p that makes the maximum, over all real values θ, of the ratio $\frac{p(\theta)}{p_0(\theta)p_L(\theta)}$ as small as possible. But the minimum over all nonnegative $q_1, \ldots, q_n$ with $q_1 + \cdots + q_n = 1$ of the maximum of $\frac{q_1}{r_1}, \ldots, \frac{q_n}{r_n}$ occurs when $\frac{q_1}{r_1} = \cdots = \frac{q_n}{r_n}$ (if they are not equal, reducing the numerator of the largest ratio, and increasing that of the smallest, will make the maximum smaller). Thus the p that makes the maximum of $\frac{p(\theta)}{p_0(\theta)p_L(\theta)}$ as small as possible is when $p(\theta) = cp_0(\theta)p_L(\theta)$, where c is chosen to make p a probability mass function, i.e., to make $p(\theta)$ sum to 1. But this is exactly the definition of the Bayesian posterior P_B in the discrete case. □

4 Minimax Likelihood Ratios

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 4.1. A 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 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, an 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 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 4.1 are similar to the min-max criterion for loss of Shannon Information (Theorem 3.5), whereas the others are dual max-min criteria. Just as the Bayesian posterior minimizes the loss of Shannon information, the Bayesian posterior is also the MLR posterior of the prior and likelihood distributions.

Theorem 4.2. *Let P_0 and L be compatible discrete or compatible absolutely continuous prior and likelihood distributions, respectively. Then the unique MLR posterior for P_0 and L is the Bayesian posterior distribution P_B.*

Proof. Immediate from [3, Theorem 5.2]. □

5 Proportional Posteriors

A criterion similar to likelihood ratios is to require that the posterior distribution P_1 reflect 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 5.1. A discrete (posterior) probability distribution P^* 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* 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 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* 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 5.2. *Let P_0 and L be 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. Immediate from [3, Theorem 5.5]. □

6 Optimal Posteriors for Weighted Prior and Likelihood Distributions

Definition 6.1. Given a prior distribution P_0 with weight $w_0 > 0$ and a likelihood distribution L with weight $w_L > 0$, the *combined weighted Shannon 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 Shannon information of the prior and likelihood always coincides with the (unweighted) combined Shannon information of the prior and likelihood. Note again that no information is attained by observing any event that is certain to occur, no matter what the distributions and weights, since $S_{P_0}(\Theta) = S_L(\Theta) = 0$. The next theorem, a special case of [4, (8)], identifies the posterior distribution that minimizes the loss of weighted Shannon information in the case the prior and likelihood distributions are compatible discrete distributions; the case for compatible absolutely continuous distributions is analogous.

Theorem 6.2. *Let P_0 and L be compatible discrete prior and likelihood distributions 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 Shannon information from the weighted prior and likelihood distributions, i.e., that minimizes, among all posterior distributions P_1,*

$$\max_A \left\{ S_{(P_0,w_0;L,w_L)}(A) - S_{P_1}(A) \right\},$$

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 [3].

References

[1] Genest, C. and Zidek, J. (1986), Combining probability distributions: a critique and an annotated bibliography, *Statist. Sci.* **1**, 114–148.

[2] Gelman, A., Rubin, D. B., Carlin, J., and Stern, H. (2004), *Bayesian data analysis*, 2nd Ed. London: Chapman & Hall.

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

[4] Hill, T. and Miller, J. (2011), How to combine independent data sets for the same quantity, *Chaos* **21**, doi:10.1063/1.3593373.