Bayesian Inference Does Not Lead You Astray... On Average

Alejandro Francetich and David M. Kreps[1]

August 2014

Alice has a six-sided die that she suspects might be loaded. Specifically, Alice assesses probability 0.9 that the die is a regular, fair die for which, on each throw, each side has probability 1/6 of appearing, independently of all other throws. But she assesses probability 0.1 that, on each throw of the die, there is probability 1/5 that the die comes up with 5 spots up, and 4/25 for each of the other five possibilities, again independently of other throws.[2]

Of course, she assesses probability one that, if she throws the die a large number of times and performs Bayesian inference after each throw, she will asymptotically be led to the truth about whether the die is fair or loaded. *In this setting, Bayesian inference leads to the truth in the (very) long run.*

But what about Bayesian inference in the short run? Suppose she throws the die once. Bayesian inference performed by Alice leads to: If the throw shows 5 spots up, she assesses (posterior) probabilities

$$\mathbf{P}[\text{die is fair}|\text{5 spots up}] = \frac{15}{17} \quad \text{and} \quad \mathbf{P}[\text{die is loaded}|\text{5 spots up}] = \frac{2}{17}\ ;$$

and if the die comes up anything other than 5, her posterior assessment is

$$\mathbf{P}[\text{die is fair}|1, 2, 3, 4, \text{or } 6 \text{ spots up}] = \frac{75}{83} \quad \text{and} \quad \mathbf{P}[\text{die is loaded}|1, 2, 3, 4, \text{or } 6 \text{ spots up}] = \frac{8}{83}\ .$$

Suppose that the die is, in fact, loaded. There is an 80% chance it will come up other than 5, in which case Alice will assess probability 8/83 = 0.0964 (approximately) that the die is loaded, less than her prior assessment 0.1. *Bayesian inference can, in the short run, lead Alice astray concerning the true state of nature*. But, assuming the die is loaded, her *expected* posterior assessment that it is loaded is

$$\frac{1}{5} \cdot \frac{2}{17} + \frac{4}{5} \cdot \frac{8}{83} = 0.10064$$

[1] We are grateful for the assistance of David Siegmund, Elie Tamer, and two anonymous referees. Specifically, David Siegmund provided the connection to Kullback-Liebler and suggested the simple proof of Proposition 1, and Elie Tamer provided the reference to the paper by I. J. Good. The financial support of ERC advanced grant 324219 and the Stanford Graduate School of Business are also gratefully acknowledged.

[2] Or, in technical terms, her overall assessment is that the infinite sequence of throws of this die is an exchangeable sequence, to which Di Finetti's theorem applies in the manner our language suggests.

(approximately), which is slightly higher than her prior. Note carefully what we are doing in this calculation: We are averaging her (ignorant-of-the-truth) posterior that the die is loaded, computing the average with the probabilities for the result of the one throw that prevail if the die is in fact loaded. And we see that, in this instance, if the die is loaded, *Bayesian inference will not lead Alice astray, on average*.[3]

This observation is not limited to Alice's problem; it is quite general. A Bayesian decision-maker (d.m.) is interested in some unobserved state of nature s and has a prior assessment about which state prevails. The d.m. observes a signal z that is (possibly) correlated with s, and forms her Bayesian posterior.

For the time being, suppose that the state s is one of a finite number of possibilities $\{s_1, \ldots, s_I\}$, and the signal z has finite support $\{z_1, \ldots, z_J\}$. Let π_i be the d.m.'s prior that $s = s_i$, and let ρ_{ij} be the likelihood that $z = z_j$ if $s = s_i$. Then, for any $i = 1, \ldots, I$, the d.m.'s posterior probability that $s = s_\ell$ if she observes $z = z_j$ is

$$\pi_{\ell|j} := \frac{\pi_\ell \rho_{\ell j}}{\sum_{i=1}^{I} \pi_i \rho_{ij}} .$$

Suppose that the true state is $s = s_\ell$, unknown to the d.m. The expected value of her posterior assessment that $s = s_\ell$, where we average over the possible signal values using their probabilities in state ℓ, is

$$\mathbf{E}\big[\pi_{\ell|z} \big| s = s_\ell\big] = \sum_{j=1}^{J} \rho_{\ell j} \pi_{\ell|z=z_j} = \sum_{j=1}^{J} \rho_{\ell j} \frac{\pi_\ell \rho_{\ell j}}{\sum_{i=1}^{I} \pi_i \rho_{ij}} .$$

Proposition 1. *For each $\ell = 1, \ldots, I$, $\mathbf{E}\big[\pi_{\ell|z} \big| s = s_\ell\big] \geq \pi_\ell$, with equality if and only if the likelihoods of the signals in state s_ℓ match precisely the marginal probabilities of the signals (so that the signal contains no useful information whatsoever).*

There is nothing surprising in this mathematical fact. One expects that Bayesian inference should work in a manner that ensures that, *in some sense*, it does not lead one astray about the "truth," no matter what the -run. But this precise instantiation of the notion that Bayesian inference does not lead one astray is not one with which we were acquainted. More to the point, we have asked a substantial number of colleagues, both in economics and statistics departments, whether they knew this specific fact, and none claimed to have known it. To be clear, not one of our colleagues was surprised by the fact; and the more ardent Bayesians in our survey complained that it was an unnatural statement from the Bayesian perspective. But, based on our survey, it does not seem well known.

[3] And if we compute her average posterior that the die is fair, if the state of nature is that it is fair, we get a number slightly larger than the prior 0.9. In this two-state case, this is obvious once you apply the well-known result that her average posterior, averaging with her current subjective probability assessment on the state of nature, must be precisely her prior.

It is not, however, unknown. I. J. Good (1965) provides among other results a theorem (his Theorem 3) that states "If an experiment has various possible experimental results, ... then the expected log-factor in favour of [the truth] ... is positive..." and which very quickly can be shown to imply the proposition. Good, in this article, attributes the result to Turing. Moreover, the result can be viewed as a straightforward corollary to the general result that the Kullback-Liebler measure of divergence (see Ghosh and Ramamoorthi, 2003, page 14) is always nonnegative.

Since the result is not well known, and since a direct proof for the finite-state, finite-signal case is very simple, we believe the result is worth restating, (re)proving, and then generalizing via Kullback-Liebler.

Proof of Proposition 1. For the finite-state, finite-signal case, let $\lambda_j := \sum_{i=1}^{I} \pi_i \rho_{ij}$; λ_j is the marginal probability of signal j . Of course, $\sum_{j=1}^{J} \lambda_j = 1$, and we can assume that $\lambda_j > 0$ for each j . (Drop from consideration any z_j whose marginal probability is zero.)

Now, write

$$\mathbf{E}\left[\pi_{\ell|z} \middle| s = s_\ell\right] = \sum_{j=1}^{J} \rho_{\ell j} \frac{\pi_\ell \rho_{\ell j}}{\lambda_j} = \pi_\ell \sum_{j=1}^{J} \frac{(\rho_{\ell j})^2}{\lambda_j} .$$

The result, then, depends on whether the last expression, without the leading factor π_ℓ , is always greater than or equal to 1.

To show that it is, proceed as follows. Let $J^* = \{j = 1, \dots, J : \rho_{\ell j} > 0\}$. Note that $\sum_{j \in J^*} \rho_{\ell j} = 1$ Then,

$$0 = \ln(1) = \ln\left(\sum_{j=1}^{J} \lambda_j\right) \geq \ln\left(\sum_{j \in J^*} \lambda_j\right),$$

with a strict inequality if J^* is a proper subset of $\{1, \dots, J\}$. And

$$\ln\left(\sum_{j \in J^*} \lambda_j\right) = \ln\left(\sum_{j \in J^*} \left[\rho_{\ell j} \cdot \frac{\lambda_j}{\rho_{\ell j}}\right]\right) \geq \sum_{j \in J^*} \rho_{\ell j} \cdot \ln\left[\frac{\lambda_j}{\rho_{\ell j}}\right],$$

where this inequality holds because the log function is concave. Therefore,

$$0 \leq \sum_{j \in J^*} \rho_{\ell j} \cdot \ln\left[\frac{\rho_{\ell j}}{\lambda_j}\right] = \sum_{j=0}^{J} \rho_{\ell j} \cdot \ln\left[\frac{\rho_{\ell j}}{\lambda_j}\right] \leq \ln\left(\sum_{j=1}^{J} \rho_{\ell j} \cdot \left[\frac{\rho_{\ell j}}{\lambda_j}\right]\right),$$

where the equality holds because the $\rho_{\ell j}$ are in the numerator, and the second inequality follows from a second use of the concavity of the log function. Now take e to the power of the first term (0) and the last term in this inequality, and you have the weak inequality. Moreover, we have a strict inequality on the first step if J^* is strictly smaller than $\{1, \ldots, J\}$. And if $J^* = \{1, \ldots, J\}$, so that $\sum_{j \in J^*} \lambda_j = 1$, we get a strict inequality in our applications of the concavity of the log function *unless* the terms $\rho_{\ell j}/\lambda_j$ are equal across the j's. But since $\sum_{j \in J^*} \rho_{\ell j} = \sum_{j \in J^*} \lambda_j = 1$ (as long as $J^* = \{1, \ldots, J\}$), the ratios are equal if and only if they are all 1, which means that we have inequality if and only if $J^* = \{1, \ldots, J\}$ and $\rho_{\ell j} = \lambda_j$ for each j; that is, if and only if the likelihoods of the signals in state s_ℓ are the same as the marginal probabilities of the various signals. ■

A stronger result

The proof just presented provides a stronger result, which is in fact the result that Good attributes to Turing:

Proposition 2. *For each* $\ell = 1, \ldots, I$, $\mathbf{E}\big[\ln(\pi_{\ell|z})\big|s = s_\ell\big] \geq \ln(\pi_\ell)$, *with equality if and only if the likelihoods of the signals in state* s_ℓ *match precisely the marginal probabilities of the signals.*

Proof. Write

$$\mathbf{E}\big[\ln(\pi_{\ell|z})\big|s = s_\ell\big] = \sum_{j=1}^{J} \rho_{\ell j} \ln\left(\frac{\pi_\ell \rho_{\ell j}}{\lambda_j}\right) = \sum_{j=1}^{J} \rho_{\ell j}\left[\ln(\pi_\ell) + \ln\left(\frac{\rho_{\ell j}}{\lambda_j}\right)\right] = \ln(\pi_\ell) + \sum_{j=1}^{J} \rho_{\ell j} \ln\left(\frac{\rho_{\ell j}}{\lambda_j}\right).$$

We then have the result if we show that $\sum_{j=1}^{J} \rho_{\ell j} \ln\left(\rho_{\ell j}/\lambda_j\right) \geq 0$. But this is precisely what we have in the last display of the previous proof. And the remark about when there is equality is also identical to the previous proof. ■

Two applications

To apply these results, enrich the setting: Suppose that the decision maker is interested in some state of nature s (drawn from a finite set S). At each date $t = 1, 2, \ldots$, she receives a signal $z(t)$ (drawn from a finite set Z_t). She begins with a prior over S, and she has a full (and accurate) set of assessments concerning the likelihoods of the various signals $z(t)$ conditional on the true state. *But beyond this, the structure of the signals in relation to one another and to the state of nature is general.* In particular, there is no presumption here that, for instance, the signals are conditionally i.i.d., conditional on the state of nature, or even that their respective supports are the same.

Despite the generality of the setting, if we let π_ℓ^t be the posterior probability assessed by the decision maker that s_ℓ is the true state, given the information gleaned from $s(1), s(2), \ldots, s(t)$, and if we let $\mathcal{P}^\ell$ denote probability conditional on $s = s_\ell$, then Propositions 1 and 2 tell us that $\{\pi_\ell^t; t = 0, 1, \ldots\}$ and $\{\ln(\pi_\ell^t); t = 0, 1, \ldots\}$ are both submartingales under $\mathcal{P}^\ell$ (for the filtration naturally generated by the sequence of signals). This has the following consequences:

1. Since $0 \leq \pi_\ell^t \leq 1$, $\{\pi_\ell^t; t = 0, 1, \ldots\}$ is a bounded submartingale, hence the sequence of posteriors converges almost surely to some π_ℓ^∞. In fact, this is nothing specially noteworthy: Under $\mathcal{P}$, the decision maker's full subjective prior, $\{\pi^t; t = 0, 1, \ldots\}$ is a bounded vector martingale and so it converges $\mathcal{P}$-a.s. As long as there is positive prior probability that the true state is s_ℓ, the same a.s. convergence holds under $\mathcal{P}^\ell$.

2. More usefully, since $-\infty \leq ln(\pi_\ell^t) \leq 1$, $\{\ln(\pi_\ell^t); t = 0, 1, \ldots\}$ is a submartingale under $\mathcal{P}^\ell$ that is bounded above and so it converges to a finite limit $\ln(\pi_\ell^\infty)$ $\mathcal{P}^\ell$-a.s. This tells us two things:

   $$\lim_{\epsilon \downarrow 0} \mathcal{P}^\ell \left[\inf_{t=1,2,\ldots,\infty} \pi_\ell^t \geq \epsilon \right] = 1,$$

   and for any $\epsilon > 0$ and any t (including $t = \infty$,

   $$\mathcal{P}^\ell \left\{ \pi_\ell^t < \epsilon \right\} \leq \frac{\ln(\pi_\ell^0)}{\ln(\epsilon)}.$$

   (The first is a direct consequence of a.s. convergence; the second follows from the submartingale inequality and, in fact, holds not only for any t but for any optional stopping time τ.) We can paraphrase these two results as: *Bayesian inference may lead you astray about the true state of nature, some of the time, but there is a finite limit how far astray it will lead you.*

   The second of these two applications is employed in Francetich and Kreps (2014).

A more general version

Proposition 1 can be generalized as follows. (We leave it to the reader to construct the general analogue to Proposition 2.) The state of nature s is drawn from some topological space S. The Borel σ-algebra on S is denoted by $\mathcal{S}$, and the d.m.'s prior on s is given by the Borel probability measure π. The signal z is drawn from a space Z with σ-algebra $\mathcal{Z}$; the (likelihood) probability measure of z given s is provided by $\mathcal{P} : \mathcal{Z} \times S \to [0, 1]$ where

a. for each $s \in S$, $\mathcal{P}(\cdot, s)$ is a probability measure on $(Z, \mathcal{Z})$, and

b. for each $A \in \mathcal{Z}$, $\mathcal{P}(A, \cdot)$ is $\mathcal{S}$-measurable.

The probability measure $\lambda : \mathcal{Z} \to [0, 1]$ on $(Z, \mathcal{Z})$ defined by $\lambda(A) := \int_S \mathcal{P}(A, s)\pi(ds)$ gives the marginal distribution of signals. Finally, a function $\pi(\cdot|\cdot) : \mathcal{S} \times Z \to [0, 1]$ is a posterior distribution of s given z if,

c. for each $z \in Z$, $\pi(\cdot|z)$ is a probability measure on $(S, \mathcal{S})$,

d. for each $B \in \mathcal{S}$, $\pi(B|\cdot)$ is $\mathcal{Z}$-measurable, and

e. for each $A \in \mathcal{Z}$ and $B \in \mathcal{S}$, $\int_B \mathcal{P}(A,s)\pi(ds) = \int_A \pi(B|z)\lambda(dz)$.

Assume that there exists a σ-finite measure μ on $(Z, \mathcal{Z})$ such that, for each $s \in S$, $\mathcal{P}(\cdot, s)$ is absolutely continuous with respect to μ. (If s and z are real-valued, and if the conditional distributions of z given s—that is, the liklihoods—have density functions, it is natural to use Lebesgue measure on the real line for μ.) Let

$$\mathcal{P}_s := \frac{d\mathcal{P}(\cdot, s)}{d\mu} : Z \to R_+$$

denote the Radon-Nikodym derivative of $\mathcal{P}(\cdot, s)$ with respect to μ.[4] Then, the Bayesian posterior distribution of s given z is given for $B \in \mathcal{S}$ by

$$\pi(B|z) = \frac{\int_B \mathcal{P}_s(z)\,\pi(ds)}{\int_S \mathcal{P}_{\hat{s}}(z)\,\pi(d\hat{s})}\ .$$

Hence, if we define the function $\beta : S \times Z \to R_+$ as

$$\beta(s,z) := \frac{\mathcal{P}_s(z)}{\int_S \mathcal{P}_{\hat{s}}(z)\,\pi(d\hat{s})}\ ,$$

$\beta(\cdot, z)$ gives, for each $z \in Z$, the Radon-Nikodym derivative of $\pi(\cdot|z)$ with respect to π.

We assume that β is measurable with respect to the product σ-algebra on $S \times Z$.[5] Then, for any $B \in \mathcal{S}$, the "average" Bayesian posterior of the d.m. over signals drawn from state $s \in B$ is, by the Fubini–Tonelli Theorem,

$$\int_{Z\times B} \beta(s,z)\,\mathcal{P}_s(z)\,(\mu \times \pi)(dz\,ds) = \int_B \left(\int_Z \beta(s,z)\,\mathcal{P}_s(z)\,\mu(dz) \right) \pi(ds).$$

Proposition 3. *For all $B \in \mathcal{S}$,*

$$\int_B \left(\int_Z \beta(s,z)\,\mathcal{P}_s(z)\,\mu(dz) \right) \pi(ds) \ \geq \ \int_B \pi(ds).$$

We will prove this by showing that, for each $s \in S$,

$$\int_Z \beta(s,z)\,\mathcal{P}_s(z)\,\mu(dz) \ \geq \ 1. \tag{1}$$

[4] The Radon-Nikodym derivative is only essentially defined; that is, two functions that differ on a μ-null set are viewed as identical. Throughout, we abuse terminology by talking about "the" Radon-Nikodym derivative.

[5] This follows if $(s,z) \to \mathcal{P}_s(z)$ is jointly measurable.

To begin, write

$$\int_Z \beta(s,z)\,\mathcal{P}_s(z)\,\mu(dz) = \int_Z e^{\ln(\beta(s,z))}\,\mathcal{P}_s(z)\,\mu(dz) \geq e^{\int_Z \ln(\beta(s,z))\mathcal{P}_s(z)\mu(dz)}\,,$$

where the last inequality follows from Jensen's inequality. So, we must show that

$$\int_Z \ln(\beta(s,z))\,\mathcal{P}_s(z)\,\mu(dz) \geq 0, \tag{2}$$

for each $s \in S$. But this follows once we show that, for each $s \in S$, the integral on the left-hand side of (2) is the *Kullback–Liebler divergence of* λ *from* $\mathcal{P}(\cdot, s)$. To explain:

> *Suppose* ν *and* ψ *are two probability measures on a measurable space* $(X, \mathcal{X})$ *such that* ν *and* ψ *are both absolutely continuous with respect to a* σ*-finite measure* η *defined on the same space. In this setting, the* Kullback–Liebler divergence of ψ from ν *is defined as*
>
> $$\mathbf{KL}(\nu, \psi) := \int_X \ln\left(\frac{d\nu/d\eta(x)}{d\psi/d\eta(x)}\right) \frac{d\nu}{d\eta}(x)\eta(dx),$$
>
> *where* $d\nu/d\eta$ *is the Radon-Nikodym derivative of* ν *with respect to* η*, and so forth. It can be shown that* $\mathbf{KL}(\nu, \psi) \geq 0$ *for all (such)* ν *and* ψ*, with equality if and only if* $d\nu/d\eta = d\psi/d\eta$ η*-a.s.*[6]

We have

$$\int_Z \ln(\beta(s,z))\,\mathcal{P}_s(z)\,\mu(dz) = \int_Z \ln\left(\frac{\mathcal{P}_s(z)}{\int_S \mathcal{P}_{\hat{s}}(z)\pi(d\hat{s})}\right)\mathcal{P}_s(z)\mu(dz). \tag{3}$$

Recall that $\mathcal{P}_s$ is the Radon-Nikodym derivative of $\mathcal{P}(\cdot, s)$ with respect to μ. For $A \in \mathcal{Z}$, the Fubini–Tonelli Theorem tells us that

$$\int_A \left(\int_S \mathcal{P}_{\hat{s}}(z)\,\pi(d\hat{s})\right)\mu(dz) = \int_S \left(\int_A \mathcal{P}_{\hat{s}}(z)\mu(dz)\right)\pi(d\hat{s}) = \int_S \mathcal{P}(A, \hat{s})\pi(d\hat{s}).$$

But by property e of conditional probabilities,

$$\int_S \mathcal{P}(A, \hat{s})\pi(d\hat{s}) = \int_A \pi(S|z)\lambda(dz),$$

[6] See Ghosh and Ramamoorthi, 2003, page 14, for details.

and since $\pi(S|z) = 1$ for all z, the right-hand side of the previous display is $\lambda(A)$. Therefore,

$$\lambda(A) = \int_A \left(\int_S \mathcal{P}_{\hat{s}}(z)\, \pi(d\hat{s}) \right) \mu(dz),$$

and so the denominator in the fraction on the right-hand side of (3) is indeed the Radon-Nikodym derivative of λ with respect to μ, completing the proof of the weak inequality in (2). Moreover, this inequality is an equation if and only if $\mathcal{P}_s(\cdot) = \int_S \mathcal{P}_{\hat{s}}(\cdot)\, \pi(d\hat{s})$ μ-a.e. Roughly put (because of μ-null sets), the inequality is an equation if, for the set of states under consideration, the signal is (a.e.) of no informational value.

References

Francetich, Alejandro, and David M. Kreps (2014). "Choosing a Good Toolkit: An Essay in Behavioral Economics," mimeo.

Ghosh, J. K., and R. V. Ramamoorthi (2003). *Bayesian Nonparametrics*. Berlin: Springer-Verlag.

Good, I. J. (1965). "A List of Properties of Bayes-Turing Factors," NSA Technical Journal, Vol. 10, No. 2, 1–6.