INNOCENZO GASPARINI INSTITUTE FOR ECONOMIC RESEARCH

Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

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

Alejandro Francetich and David Kreps

Working Paper n. 514

This Version: September, 2014

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
http://www.igier.unibocconi.it

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

Alejandro Francetich and David M. Kreps[1]

September 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

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

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 De 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]

In fact, we can get a stronger result. The expectation of the *log* of her posterior that the die is loaded, computed in the same way (using the probability distribution in force if in fact the die is loaded), is

$$\frac{1}{5} \cdot \ln\left(\frac{2}{17}\right) + \frac{4}{5} \cdot \ln\left(\frac{8}{83}\right) = -2.29953 > -2.3026 = \ln(0.1).$$

The expectation, taken in this fashion, of the (natural) log of her posterior is greater than the log of her prior.

This observation is not limited to Alice's problem. Suppose a Bayesian decision maker is interested in the probability of some state of nature S. Specifically, she is interested in the probability that $\{S \in B\}$, for some event B such that $\pi := \mathbf{P}(S \in B) > 0$, where $\mathbf{P}$ denotes probability according to her prior. She obtains information about S in the form of a signal Z that can take on finitely many values $z_1, \ldots, z_J$. Let ρ_j and ρ'_j be the likelihoods that $\{Z = z_j\}$ if $S \in B$ and if $S \in B^C$ (the complement of B), respectively. And let $\lambda_j := \pi\rho_j + (1-\pi)\rho'_j$; λ_j is, of course, the marginal probability of seeing signal z_j. Discard from further consideration any signal that is "impossible"; i.e., assume $\lambda_j > 0$ for all j. Then, if she observes that $Z = z_j$, her posterior assessment that $\{S \in B\}$, by Bayes rule, is

$$\pi_{B|Z=z_j} := \mathbf{P}\big[S \in B \big| Z = z_j\big] = \frac{\pi\rho_j}{\lambda_j}.$$

Let $J^* \subseteq \{1, \ldots, J\}$ be the subset of possible values for Z that have strictly positive likelihood if $S \in B$. Supposing that (unknown to the decision maker) S is indeed in B, the expectation of the log of her posterior assessment that $\{S \in B\}$, given the range of possible signals, is

$$\mathbf{E}^B\big[\ln(\pi_{B|Z})\big] = \sum_{j \in J^*} \rho_j \ln\left(\pi_{B|Z=z_j}\right) = \sum_{j \in J^*} \rho_j \ln\left(\frac{\pi\rho_j}{\lambda_j}\right),$$

where the expectation is over the possible values of Z (hence we drop "$= z_j$" in $\pi_{B|Z}$), and the superscript B on the expectation operator signifies that the expectation is with respect to the probability distribution of Z that holds when $S \in B$.[4]

[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.

[4] If z_j is such that $\rho_j = 0$, the term $\rho_j \cdot \ln(\pi\rho_j/\lambda_j)$ is zero times minus infinity. We exclude such j in the sum that forms the expectation of the log of the posterior; you can include them in this sum if, by convention, you assume that the product is zero. Note that $\lim_{x \downarrow 0} x \ln(kx) = 0$, so this convention is appropriate in the limit.

Proposition 1. $\mathbf{E}^B\left[ln(\pi_{B|Z})\right] \geq ln(\pi)$, *with equality if and only if* $\rho_j = \rho'_j$ *for all* j *(that is, if the signals are uninformative about whether* $S \in B$ *or not).*

Corollary 1. $\mathbf{E}^B\left[\pi_{B|Z}\right] \geq \pi$, *with equality if and only if* $\rho_j = \rho'_j$ *for all* j.

There is nothing surprising in these mathematical facts. 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.

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..." This, essentially, is 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 is very simple, we believe the result is worth restating and (re)proving. Here is the proof:

Of course, $\sum_j \lambda_j = 1$, and so

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

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

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

where the inequality holds by Jensen's inequality (the log function is concave). Switching numerator and denominator inside the log changes the sign of the inequality, and so

$$0 \leq \sum_{j\in J^*} \rho_j \cdot \ln\left[\frac{\rho_j}{\lambda_j}\right] = \sum_{j\in J^*} \rho_j \cdot \ln\left[\frac{\rho_j}{\lambda_j}\right].$$

Add $\ln(\pi)$ to both sides to get

$$\ln(\pi) \leq \ln(\pi) + \sum_{j \in J^*} \rho_j \cdot \ln\left[\frac{\rho_j}{\lambda_j}\right] = \sum_{j \in J^*} \rho_j \cdot \ln\left[\frac{\pi\rho_j}{\lambda_j}\right].$$

Note that the inequality is strict if either J^* is a proper subset of $\{1, \ldots, J\}$ or if the ratios λ_j/ρ_j for $j \in J^*$ are not all the same, which is true if and only if each $\rho_j = \rho'_j$. This proves Proposition 1; the corollary is immediate from another application of Jensen's inequality.

Two applications

To apply these results, enrich the setting: Suppose that a Bayesian decision maker is interested in whether some state of nature S belongs to an event B (with positive prior probability). At each date $t = 1, 2, \ldots$, she receives a signal $Z(t)$ (drawn from a finite set $\mathcal{Z}(t)$). 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^t be the posterior probability assessed by the decision maker that $S \in B$, given the information gleaned from $Z(1), Z(2), \ldots, Z(t)$, and if we let $\mathbf{P}^B$ denote probability conditional on $\{S \in B\}$, then Proposition 1 and the Corollary tell us that $\{\pi^t; t = 0, 1, \ldots\}$ and $\{\ln(\pi^t); t = 0, 1, \ldots\}$ are both submartingales under $\mathbf{P}^B$ (for the filtration naturally generated by the sequence of signals), where π^0 is the prior, or $\mathbf{P}\big(\{S \in B\}\big)$. This has the following consequences:

1. Since $0 \leq \pi^t \leq 1$, $\{\pi^t; t = 0, 1, \ldots\}$ is a bounded submartingale, hence the sequence of posteriors converges almost surely to some π^∞. In fact, this is nothing specially noteworthy: Under $\mathbf{P}$, the decision maker's full subjective prior, $\{\pi^t; t = 0, 1, \ldots\}$ is a bounded martingale and so it converges $\mathbf{P}$-a.s. Since (we assume) $\mathbf{P}(\{S \in B\}) > 0$, the same a.s. convergence holds under $\mathbf{P}^B$.

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

$$\lim_{\epsilon \downarrow 0} \mathbf{P}^B\left[\inf_{t=1,2,\ldots,\infty} \pi^t \geq \epsilon\right] = 1,$$

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

$$\mathbf{P}^B\left[\{\pi^t < \epsilon\}\right] \leq \frac{\ln\left(\mathbf{P}\left[\{S \in B\}\right]\right)}{\ln(\epsilon)}.$$

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 one can bound the probability that it leads you far astray.*

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

Signals with continuous density functions

While the proposition is stated for general states of nature S, it assumes that the signal Z takes on one of a finite number of values. In many models of interest, Z will be a continuous random variable. Suppose that, given $S \in B$, the signal Z has continuous density function $g(z)$. Suppose that on the complement of $S \in B$, it has continuous density function $g'(z)$. Letting π continue to be the prior probability of $\{S \in B\}$, the "natural"[5] application of Bayes' Rule leads the decision maker, having observed $Z = z$, to assess a posterior for $\{S \in B\}$ of

$$\pi_{B|Z=z} = \frac{\pi g(z)}{\pi g(z) + (1-\pi)g'(z)}.$$

And the expectation of the log of this posterior, computed assuming $S \in B$, is

$$\mathbf{E}^B\left[\ln\left(\pi_{B|Z}\right)\right] = \int_R \ln\left(\frac{\pi g(z)}{\pi g(z) + (1-\pi)g'(z)}\right) g(z)dz,$$

where we conventionally assume that the integrand is zero when $g(z)$ is zero, or we integrate only over those z for which $g(z) > 0$.

Corollary 2. $\mathbf{E}^B\left[ln\left(\pi_{B|Z}\right)\right] \geq ln(\pi)$, *and* $\mathbf{E}^B\left[\pi_{B|Z}\right] \geq \pi$, *with equality if and only if* $g \equiv g'$.

The proof essentially enlists the definition of the integral. Let R^* be the subset of R on which $g(z) > 0$ and, for each $n = 1, 2, \ldots$, partition R^* into n intervals whose diameter goes to zero, except possibly for the left-most and right-most of these intervals. For each $z \in R^*$, let $I^n(z)$ be the interval in the nth partition that contains z. For each partition, construct an "approximate" signal Z^n that reveals which interval contains Z. Apply Proposition 1 to this nth-approximation. Then use the continuity of $g(z)$ and $g'(z)$ to show that, as $n \to \infty$, $\mathbf{E}^B[\ln(\pi_{B|Z^n})]$ has limit $\mathbf{E}^B[\ln(\pi_{B|Z})]$.

What if the state S is a continuous random variable with, say, a prior given by a continuous density f? Note that we have made no assumptions about the nature of S, but we are looking at a "discrete" event $\{S \in B\}$ with strictly positive prior probability. If you want a result that invokes the prior density

[5] That is, this rule, applied for all s, gives a version of the conditional probability.

f directly (at, say, a specific value s for which $f(s) > 0$), simply take a sequence of intervals B_n that surround s and have vanishing diameter, and pass to the limit.

A very general version

We close with a very general and abstract version of this sort of result. 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 decision maker'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.[6] Then, the Bayesian posterior distribution of s given z is given for $B \in \mathcal{S}$ by

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

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

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

[6] 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.

$\beta(\cdot, z)$ 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 $S \times Z$.[7] Then, for any $B \in \mathcal{S}$, the "average" Bayesian posterior of the decision maker 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 $s \in S$,

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

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*
>
> $$\mathbf{KL}(\nu, \psi) := \int_X \ln\left(\frac{d\nu/d\eta(x)}{d\psi/d\eta(x)} \right) \frac{d\nu}{d\eta}(x)\eta(dx),$$

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

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.[8]

We have

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

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

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

But by property e of conditional probabilities,

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

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

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

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 $\mathcal{P}_s(\cdot) = \int_S \mathcal{P}_{s'}(\cdot)\,\pi(ds')$ 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.

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