LA-UR-13-25883

Approved for public release; distribution is unlimited.

Title: Brittleness and Bayesian Inference

Author(s): Wallstrom, Timothy C.

Intended for: arXiv
Report

Issued: 2013-08-15 (rev.1)

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Brittleness and Bayesian inference*

Timothy C. Wallstrom[†]

Abstract. In a recent report, “Bayesian Brittleness: Why no Bayesian model is ‘good enough,’” (arXiv, 1304.6772v1) the authors prove rigorously that under a seemingly natural class of priors Π , the posterior expectation of a function $\Phi(\theta)$ can achieve essentially any value that $\Phi(\theta)$ alone can achieve, so that Bayesian inference appears to have no robustness whatsoever. We explain this puzzling result, and show why it does not imply a breakdown of Bayesian inference. The problem is that the priors leading to the extreme results depend on the observed data, which means they are not valid priors. The corresponding posteriors are also invalid, so the extreme variation in these posteriors does not imply that Bayesian inference is brittle. The report, however, provides a new framework which may be used for any assumed class of priors Π , and which we explain in detail. This framework may be a useful new tool for deriving quantitative robustness results.

1. Introduction. A central challenge in Bayesian inference is to determine the prior $\pi(\theta)$, since there may be a range of choices which seem more or less equally plausible. The idea of robust Bayesian inference is to consider the set of plausible priors Π , and then to look at how inferences vary across this set. If the smallest and largest inferences are similar, then the inference is “robust”: it does not depend strongly on any uncertainty in the input prior. If, on the other hand, these inferences are widely different, one concludes that the inference is “fragile,” or “brittle.”

The problem of Bayesian robustness is studied in an unpublished but widely circulated recent report, “Bayesian Brittleness: Why no Bayesian model is ‘good enough,’” by Owhadi, Scovel, and Sullivan [3]. The authors introduce a novel method for constructing Π , which provides a class of priors that seem to be very similar to each other. For example, the priors in Π might all imply the same moment distributions for the output variable X , up to some arbitrarily large order k , or might be concentrated on distributions that are arbitrarily close to some parametric model class.

The authors then prove rigorously that under their set of priors Π , Bayesian inference has no robustness whatsoever! If $\Phi(\theta)$ is some output of interest, and $E_\pi(\Phi|d)$ is the posterior expectation of Φ after seeing data d , then for different choices of $\pi \in \Pi$, and under “extremely weak conditions,” the posterior expectation can have any value between the minimum and maximum of Φ under variations in θ .

The purpose of the present paper is to explain this puzzling result. The explanation, in brief, is that the priors depend on the observed data, as will be clear when we look at specific examples. If the prior depends on the data, it is possible to put the posterior mass anywhere. Thus, one can put the mass on values of θ for which $\Phi(\theta)$ is large, or small, or anything in between.

It is true that the space Π is defined without reference to the data. It turns out, however, that Π is so large that it contains π that are “pre-adapted” to any possible dataset. In

*This work supported by Quantification Methods, a part of NNSA’s Advanced Simulation and Computing Program, and from the Department of Energy under contract DE-AC52-06NA25396.

[†]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, tcw@lanl.gov.

particular, for any dataset, we can find a “prior” for which the posterior is only nonzero where we want it to be nonzero. In this way, we can make the posterior expectation whatever we want. But since the priors depend on the data, these posteriors are not meaningful, and provide no evidence for fragility in Bayesian inference.

It is well-known that problems can arise from an excessively large Π . Here, for example, is a discussion from a 1994 review paper by Berger [1]:

“The size of Π should be appropriate, in the sense of being a reasonable reflection of prior uncertainty. If Π is too small, one might fear being erroneously led to a conclusion of robustness. If Π is too large, in the sense of containing many prior distributions that are clearly unreasonable, then one might conclude that robustness is lacking when, in fact, a reasonable Π would imply robustness. For detecting this latter problem, it is useful to determine the $\pi \in \Pi$ at which [the infimum or supremum of the posterior mean of Φ] is attained, and judge if such a π is reasonable. If not, one should try to refine Π to eliminate such π .” [1, p. 24].

In the remainder of the paper, I will first describe the procedure for defining Π (§2), and for finding priors in Π that lead to large or small values of the posterior expectation of Φ (§3). This discussion will make it intuitively clear how the results of [3] are obtained, and why they are mathematically correct. Following Berger’s advice, I then examine the π that were used in [3] to generate extreme results, and point out how they depend on the data (§4). In the discussion (§5), I suggest a way in which the framework of [3] might be used, with a different set of assumptions about Π , as a new tool to study robustness. Finally, in the Appendix, I relate my assumptions and notation to those of [3].

2. Construction of Π . We now describe how Π , the space of plausible priors π , is constructed. Our treatment is formal, but a rigorous treatment in a more general setting can be found in [3].

Let Θ be the parameter space, where each $\theta \in \Theta$ is a measure on $\mathcal{X}$. That is, we take θ to be the measure itself, rather than some set of parameters characterizing the measure. If the data D is distributed according to θ , then

$$p(D \in dx|\theta) = \theta(dx).$$

Priors are then probability measures on $\mathcal{M}(\mathcal{X})$, the space of measures on $\mathcal{X}$.

To describe Π , we first need to introduce the space $\mathcal{Q}$, which we call the *base space*, and the space M , which we call the *bundle space*. For each $q \in \mathcal{Q}$, let Θ_q be a subset of Θ . Let

$$M = \{(q, \theta) : q \in \mathcal{Q}, \theta \in \Theta_q\}, \quad (2.1)$$

and define, for $m \in M$, the *projection operator* Ψ as the function taking m to q . The *fiber over q* is the set $F_q \equiv \Psi^{-1}(q)$, which may be identified with Θ_q under the mapping $(q, \theta) \mapsto q$. A *section* ψ is a mapping from $\mathcal{Q}$ to M , such that $\psi(q) \in F_q$, or equivalently, such that $\Psi \circ \psi(q) = q$. If ψ is a section, we frequently use θ_q to denote the second component of $\psi(q)$, as in $\psi(q) = (q, \theta_q)$. The setup is shown in Figure 2.1.

The idea of using a base space is that q characterizes some feature of the measures in Θ_q , so that all of the measures in Θ_q are similar in some sense. The two main examples are the following:

Figure 2.1. Each element $q \in \mathcal{Q}$ has a fiber F_q above it in M . The projection map takes any point $m' \in M$ to $\Psi(m') \in \mathcal{Q}$, and the section ψ takes any point $q'' \in \mathcal{Q}$ to $\psi(q'') = (q'', \theta_{q''}) \in M$. The space $\mathcal{Q}$ may or may not have a natural embedding in M .

- *Moment space.* Here, $\mathcal{Q}$ is the vector of moments to order k . Thus, the projection operator is

$$\Psi(\theta) = q = (\mathbf{E}_\theta D, \mathbf{E}_\theta D^2, \dots, \mathbf{E}_\theta D^k). \quad (2.2)$$

Θ_q is the space of all measures with moments $q = (q_1, \dots, q_k)$, which we might regard as similar in some sense, particularly if k is large.

- *Model class space.* Here, $\mathcal{Q}$ is the parameter space for some finite-dimensional model class. Thus, for example, $q = (a, b)$ if the model class is the space of Beta distributions, $B(x; a, b)$.¹ We define θ_q^0 to be the measure corresponding to the parameter q ; the mapping $\psi_0 : q \mapsto (q, \theta_q^0)$ may be regarded as a distinguished section. Θ_q^α can now be defined as the set of all measures θ with $d(\theta_q^0, \theta) < \alpha$, where $d(\cdot, \cdot)$ is some metric on the space of measures. Again, the idea is that the elements of Θ_q are similar, particularly if α is small.

We can now define the space Π . Suppose we are given a measure $\mathbb{Q}$ on $\mathcal{Q}$. The measures π are defined by drawing q randomly according to $\mathbb{Q}$, and then drawing θ randomly from Θ_q , according to some new random measure on the fiber determined by π . In symbols,

$$\pi(A) = \int \pi(A|q) \mathbb{Q}(dq), \quad (2.3)$$

where $A \subset \mathcal{M}(\mathcal{X})$. The measures $\pi \in \Pi$ correspond to all of the different ways of choosing the fiber measures $\pi(A|q)$. The space Π is therefore very large. One might hope, however, that if the θ in Θ_q are very similar, then the π 's might also be similar, and that Π might capture the uncertainty in the value of the prior. An argument for specifying priors in this way can be found in the last paragraph of [3, Section 2.2].

For purposes of computing extrema, it turns out that we can focus on a particular subset of Π , in which the $\pi(\cdot|q)$ are concentrated on a single parameter, i.e.,

$$\pi(\cdot|q) = \delta_{\theta_q}(\cdot), \quad (2.4)$$

¹Note that we are using the word “parameter” in two senses: q is the usual finite-dimensional parameterization of the model class, and θ is the “parameter” in the space of infinite-dimensional distributions. The context should make the meaning clear.

where $\delta_\theta(\cdot)$ is the Dirac point mass at θ . (This fact is intuitively reasonable, and is essentially the content of [3, Theorem 4.11].) Any such prior is conveniently associated with the section $\psi : q \mapsto (q, \theta_q)$. This subset of priors, which might be denoted Π_ψ , is still enormous, because each θ_q can be arbitrarily chosen from Θ_q , which is generally infinite-dimensional. When π corresponds to a section ψ , we denote it by π_ψ . By Eq. 2.3, $\pi_\psi(A) = \int \delta_{\theta_q}(A) \mathbb{Q}(dq)$.

3. Finding priors that give extrema. We begin by formulating our problem more precisely. Let $\Phi(\theta)$ be some observable of interest. We are interested in the posterior mean of Φ , conditioned on the data $d = (d_1, \dots, d_n)$ lying in the set B :

$$\mathbf{E}_\pi(\Phi|B) = \int \Phi(\theta) \pi(d\theta|B) \quad (3.1)$$

$$= \frac{\int \Phi(\theta) p(B|\theta) \pi(d\theta)}{p_\pi(B)}, \quad (3.2)$$

where $p_\pi(B) = \int p(B|\theta) \pi(d\theta)$. We condition on B , rather than on the data d , to avoid problems with sets of measure zero. Specifically, we set

$$B_{\epsilon,i} \equiv \{d : \rho(d, d_i) < \epsilon\}, \quad (3.3)$$

where $\rho(\cdot, \cdot)$ is some distance metric on $\mathcal{D}$, and take

$$B = B_{\epsilon,1} \times \dots \times B_{\epsilon,n}.$$

We reserve the right to take $\epsilon > 0$ as small as desired.

To prove the brittleness theorems, we want to show, e.g., that

$$\mathbf{E}_\pi(\Phi|B) > \lambda,$$

where λ is some threshold that is arbitrarily close to the supremum of $\Phi(\theta)$. In Theorem 5.12 of [3] (the “Main Brittleness Theorem”), it is shown how to prove this result using essentially the following two Assumptions:

1. On almost all fibers (with respect to $\mathbb{Q}$), there exists θ such that $p(B|\theta) = 0$.
2. There is a set of fibers with positive $\mathbb{Q}$ measure such that each contains a θ with $p(B|\theta) > 0$ and $\Phi(\theta) > \lambda$.

To prove the theorems, we pick a section prior π_ψ , with $\psi(q) = (q, \theta_q)$, as follows. If q satisfies Assumption 2, let $\theta_q = \theta$, where $\Phi(\theta) > \lambda$. Otherwise, let $\theta_q = \theta$, where $p(B|\theta) = 0$.

Assumption 1 enables us to suppress the contribution of any q at will. If there is no $\theta \in \Theta_q$ such that $\Phi(\theta) > \lambda$, we just pick a θ_q such that $p(B|\theta_q) = 0$, and this fiber then drops out of the posterior averaging, as shown by Eq. 3.2. Assumption 2, on the other hand, enables us to ensure that desirable fibers, i.e., those with some θ such that $\Phi(\theta) > \lambda$, *do* occur in the average with positive probability. Since the average will be over only those fibers, the posterior mean will be greater than λ .

It is not at all obvious, however, why these Assumptions should be true, and in fact, their validity depends on the problem. The best way to understand their significance is to look at particular examples.

4. Examples. We now consider the two examples mentioned in Section 2. It will simplify the discussion to assume that $\mathcal{X} = [0, 1]$; we let dx denote Lebesgue measure on $\mathcal{X}$. It will be obvious, however, that this assumption in no way affects the generality of our conclusions.

4.1. Moment space. We assume first that $k = 1$ and $n = 2$, and discuss generalizations below. Our construction of π follows [3, Example 5.15].

Suppose, then, that $\mathcal{X} = [0, 1]$, Θ is the space of all measures on $\mathcal{X}$, and $\Phi(\theta) = \mathbf{E}_\theta(D)$. The base space $\mathcal{Q} = [0, 1]$, so that the fiber above q is the set of all measures θ with $\mathbf{E}_\theta(D) = q$. The maximum possible value of $\Phi(\theta)$ is 1, so for any $\delta > 0$, we want to find a prior π for which $\mathbf{E}_\pi(\Phi|B) > 1 - \delta$.

Let $d = (d_1, d_2)$, and assume that $d_1 \neq d_2$, which is true with probability one if the true measure is absolutely continuous with respect to dx . Choose $\epsilon > 0$ small enough that $B_{\epsilon,1} \cap B_{\epsilon,2} = \emptyset$. Let

$$\mu_d = \frac{1}{2}(\delta_{d_1} + \delta_{d_2})$$

be the empirical distribution of the sample. Define $\pi = \pi_\psi$ to be the section prior such that

$$\theta_q = \begin{cases} \delta_q & \text{(Case 1: } q \leq 1 - \delta) \\ (1 - \epsilon_q)\delta_{q'} + \epsilon_q\mu_d & \text{(Case 2: } q > 1 - \delta), \end{cases} \quad (4.1)$$

where, in Case 2, q' is chosen so that $\mathbf{E}_{\theta_q}X = q$. By making ϵ_q suitably small, we can ensure that $0 \leq q' \leq 1$.

Then, for q in Case 1,

$$\begin{aligned} p(B|\theta_q) &= p(B_{\epsilon,1}|\theta_q) p(B_{\epsilon,2}|\theta_q) \\ &= 0, \end{aligned}$$

because q cannot be in two disjoint sets simultaneously. On the other hand, for Case 2,

$$p(B|\theta_q) = \frac{1}{4}\epsilon_q^2 > 0,$$

so these values *do* enter into the posterior mean. Since $\Phi(q) = q > 1 - \delta$ for these q , and since these are the only values entering into the posterior mean, the same inequality holds for the posterior mean.

The key point is that in Case 2, we have had to contaminate the prior for $q > 1 - \delta$ with the empirical distribution, in order to ensure that $p(B|\theta_q) > 0$. To do so required knowledge of the data, so π_ψ is not a valid prior.

Incidentally, these “priors” are peculiar in other ways as well. In order to ensure that the d_i are distinct, we have had to assume that the true measure is absolutely continuous with respect to Lebesgue measure. But that assumption implies that we know, *a priori*, that there are no Dirac δ ’s in the distribution, and if that is the case, it seems inconsistent to adopt a prior that is a finite sum of such δ ’s.

The authors show how to carry out the analogous construction for arbitrary k and n . For $k > 1$, any prescribed set of (attainable) moments can be achieved using a normalized sum of k Dirac δ ’s. As long as there are at least $k + 1$ distinct datapoints, Assumption 1 is satisfied, because k δ ’s cannot lie in every one of $k + 1$ disjoint sets. Assumption 2 is again established by contaminating a Dirac measure with the empirical distribution.

Figure 4.1. The black line is the density of θ_q^0 . The red line is the modified density, $\theta_{0,q}^c$, given that the data is $d = (0.07, 0.17, 0.32, 0.52, 0.64, 0.71, 0.82)$, and $\epsilon = 0.005$. We can make the total variation distance between these two measures smaller than any $\alpha > 0$ by making ϵ small enough. But it is hard to see how anyone would have put any prior mass on the second measure, unless they had first seen the data.

4.2. Model class space. Let $\mathcal{Q}$ be the parameter space of some model class, and let $\Phi(\theta) = \mathbf{E}_\theta(D)$. In this problem, $\mathcal{Q}$ might be regarded as the original prior in the parametric model. As above, let

$$\Theta_q^\alpha = \{\theta : d_{TV}(\theta, \theta_q^0) < \alpha\}, \quad (4.2)$$

where $\alpha > 0$, and where $d_{TV}(\cdot, \cdot)$ is the total variation distance:

$$d_{TV}(\theta_1, \theta_2) = \sup_A |\theta_1(A) - \theta_2(A)|. \quad (4.3)$$

$$= \frac{1}{2} \int |\theta_1(x) - \theta_2(x)| dx, \quad (4.4)$$

where A is any measurable subset of $\mathcal{X}$, and the second expression applies when the θ_i have densities with respect to dx . We will further assume that all elements of the model class are strictly positive on $\mathcal{X}$. This assumption is adopted for convenience, but is often valid for parametric families. Such will be the case, for example, if our model class is the set of Beta distributions on $[0, 1]$.

We wish to find $\pi \in \Pi$ such that $\mathbf{E}_\pi(\Phi|B) > 1 - \delta$. Following the proof of Theorem 6.10 in [3], we can do this by defining $\pi = \pi_\psi$, where $\psi : q \rightarrow (q, \theta_q)$, and

$$\theta_q = \begin{cases} \theta_q^{0,c} & (\text{Case 1: } \Phi(\theta_q^0) \leq 1 - \delta) \\ \theta_q^0 & (\text{Case 2: } \Phi(\theta_q^0) > 1 - \delta). \end{cases}$$

Here,

$$\theta_q^{0,c}(A) \equiv \theta_q^0(A|B^c) = \frac{\theta_q^0(A \cap B^c)}{\theta_q^0(B^c)},$$

where B^c is the complement of B .

In Case 1,

$$p(B|\theta_q) = \theta_q^{0,c}(B) = 0,$$

where the last equality is immediate from the definition of $\theta_q^{0,c}$. These values of q , therefore, will not figure into the posterior mean. In Case 2, $\Phi(\theta_q) > 1 - \delta$ by definition, and our

positivity assumption implies that $p(B|\theta_q) > 0$. Thus, the posterior mean of the functional will be greater than $1 - \delta$ also. (If we had not made a positivity assumption, we could have contaminated θ_q with a small amount of the empirical distribution, as before.)

It remains to show that for all q , θ_q is in Θ_q^α , which is to say, within a distance α of θ_q^0 . This is only a concern when $\theta_q = \theta_q^{0,c}$, since otherwise $\theta_q = \theta_q^0$, and the distance is zero. An algebraic proof is simple, but we illustrate the situation graphically in Figure 4.1. The modified distribution has been set to zero in the regions where the data actually occurred. There are two contributions to the total variation, corresponding to the regions away from the data and the regions near the data. Both of these scale with $\sum_{i=1}^n \theta_q^0(B_{\epsilon,i})$, and can be made arbitrarily small by making ϵ arbitrarily small.

Figure 4.1 also illustrates how the prior depends on the data. The notches in the distribution are unnatural, and you would not know where to put them until you saw the data. The prior used in this example is therefore not a valid prior.

5. Discussion. In the usual approaches to Bayesian robustness, we restrict our attention to a finite-dimensional parameterized model class. To assess robustness, we consider variation in the priors that might be assigned to the parameters in the class. In so doing, however, as the authors of [3] point out, we are making a very strong assumption. The parametrically defined models lie on a finite-dimensional submanifold of the infinite-dimensional space of all possible models. Even if we vary over the priors on the finite-dimensional parameter space, we are not considering any variation on the parameters that could be introduced to describe the infinite-dimensional complementary space. (Let us call these “co-parameters.”) In effect, we are setting an infinite set of co-parameters equal to zero, which seems like a very strong and unrealistic assumption, particularly if we are interested in robustness. And in practice, whenever we study robustness in real problems, we know that the true mechanism of uncertainty will not lie in our model space; these co-parameters are *not* equal to zero. This is particularly obvious in applications to complex systems or machine learning, for example.

The paper under consideration is an effort to address this problem. The approach is to assume, instead, that what is known is finite, and that what is unknown is infinite. Specifically, the values of q are handled conventionally, in that a class of priors $\mathbb{Q} \in \mathfrak{Q}$ is considered. In addition, however, fibers are constructed over each q , and an additional measure $\pi(\cdot|q)$ is defined on Θ_q , in order to represent prior uncertainty about the co-parameters.

As we have seen, however, if we allow $\pi(\cdot|q)$ to be chosen arbitrarily, we wind up allowing priors that are not reasonable representations of prior belief. The key problem is that this freedom can be used, not to represent uncertainty about the co-parameters, but to represent states of extreme knowledge about the co-parameters, which may be both highly detailed and implausible. Thus, in the moment example, if we take the measure on the fiber above q to be δ_q , we have effectively assumed that we know all the moments, from 1 to ∞ . This is not a state of ignorance about the co-parameters. And, as we have seen, we can construct what might be called “psychic” priors, that reflect knowledge of exactly where the data will appear in the future.

In order to carry out the idea of [3], it seems that we would want $\pi(\cdot|q)$ to represent *ignorance* of the co-parameters. Such ignorance could be represented, for example, by adopting a Dirichlet prior, $D(\theta_q^\alpha)$, on the fiber, with $\theta_q^\alpha = \alpha\theta_q^0$; for details, see [2]. As $\alpha \rightarrow \infty$, the

measures $D(\theta_q^\alpha)$ will become more and more concentrated on measures close to θ_q^0 . In general, we would represent our uncertainty on q with $\mathbb{Q}$, for a variety of measures $\mathbb{Q} \in \mathfrak{Q}$, and our uncertainty about the fiber by a diffuse prior (or set of such priors) on that fiber. In this way, the framework constructed in [3] might be useful in relaxing the unrealistic assumption that the true model lies in the model class, and in determining how this impacts the variability in posterior inferences, or in other words, their robustness.

Acknowledgments. I thank David Higdon for bringing this interesting paper to my attention. I also thank the authors for considering my ideas, and providing me with their feedback. I gratefully acknowledge support from Quantification Methods, a part of NNSA's Advanced Simulation and Computing Program, and from the Department of Energy under contract DE-AC52-06NA25396.

Appendix: Comparison of notation and results. I provide some guidance for the reader wishing to compare the present setup with the original setup in [3]. In the present setup, I have changed the notation somewhat, so as to simplify the presentation and to emphasize the connections with the central concepts of Bayesian inference. I have also suppressed much of the generality, in order to isolate the essential features leading to extreme posteriors. I first provide a general framework, which I then specialize to the two different setups. The reader should be aware, however, that [3] is an arXiv preprint, and the notation in that paper may change in future versions.

I work in the following general setting. A single datapoint d is a realization of the random variable D , and takes values in $B \subset \mathcal{D}$ with probability

$$p(B|\theta),$$

where $p(\cdot|\theta)$ is a *statistical model*, and θ , the *parameter*, takes values in Θ . When there are several samples, so that

$$d = (d_1, \dots, d_n),$$

we assume that the d_i are independent, so that if $B = B_1 \times \dots \times B_n$, then

$$p(d \in B|\theta) = \prod p(d_i \in B_i|\theta).$$

There may be some uncertainty about the correct statistical model; we indicate a particular choice with the notation $p_{\mathbb{D}}(\cdot|\theta)$, where $\mathbb{D}$ belongs to some model space $\mathfrak{D}$. Given a prior π on Θ , we define the joint distribution,

$$p_{\pi, \mathbb{D}}(A, B) = \int_A p_{\mathbb{D}}(B|\theta) \pi(d\theta),$$

the marginal data distribution

$$p_{\pi, \mathbb{D}}(B) = \int_{\Theta} p_{\mathbb{D}}(d\theta, B),$$

and the posterior

$$\pi_{\pi, \mathbb{D}}(A|B) = \frac{p_{\pi, \mathbb{D}}(A, B)}{p_{\pi, \mathbb{D}}(B)},$$

where $A \subset \Theta$, and $B \subset \mathcal{D}$. The posterior expectation of $\Phi(\theta)$, given that $d \in B$, is

$$\mathbf{E}_{\pi, \mathbb{D}}(\Phi|B) = \int_{\Theta} \Phi(\theta) \pi_{\pi, \mathbb{D}}(d\theta|B).$$

In robust Bayesian inference, we define a space Π of possible priors π , and a space $\mathfrak{D}$ of possible models $\mathbb{D}$ [1]. We then compute the supremum and infimum of the posterior expectations of Φ :

$$\overline{\Phi}_B = \sup_{\pi \in \Pi, \mathbb{D} \in \mathfrak{D}} \mathbf{E}_{\pi, \mathbb{D}}(\Phi|B) \quad \text{and} \quad \underline{\Phi}_B = \inf_{\pi \in \Pi, \mathbb{D} \in \mathfrak{D}} \mathbf{E}_{\pi, \mathbb{D}}(\Phi|B).$$

If $\overline{\Phi}_B - \underline{\Phi}_B$ is small, the inference is robust.

Applying this now to the setup in [3], the data are realizations of

$$D = (X, f(X) + \epsilon) = (X, Y),$$

where $X \in \mathcal{X}$ is drawn randomly from a distribution μ , and ϵ from a distribution ν , say. The quantities μ , ν , and f are all unknown. The data are thus generated by a statistical model of the form $p_{\mathbb{D}}(B|\theta)$, where $d = (x, y)$, $\theta = (f, \mu)$, and $\mathbb{D}$ embodies the distribution ν of ϵ . The model can be parsed as follows:

$$p_{\mathbb{D}}(dx \times dy|\theta) = p_{\mathbb{D}}(dy|f, x) p(dx|\mu),$$

where

$$y|f, x \sim f(x) + \epsilon.$$

In the present setup, the data are realizations of D , where $D \in \mathcal{X}$ is distributed according to θ :

$$p(D \in B|\theta) = \theta(B).$$

Thus, in comparison to [3], I take f to be the identity, and $\mathcal{D} = \mathcal{X}$. I do not consider both X and Y ; I only consider $Y = X + \epsilon$ (which I identify as the data D). I do not differentiate between uncertainty due to μ and that due to ϵ ; their net effect is subsumed into θ . In particular, I do not explicitly model or extremize over the space $\mathfrak{D}$. I also have a less general construction of Π , in that I assume only one measure, which I denote $\mathbb{Q}$, on the base space $\mathcal{Q}$. In [3], the authors consider a set $\mathfrak{Q}$ of measures $\mathbb{Q}$ on $\mathcal{Q}$. For each $\mathbb{Q} \in \mathfrak{Q}$, they construct a set of priors, using the construction above; the space Π is then the union of such sets of priors over all $\mathbb{Q} \in \mathfrak{Q}$. This union is written $\Psi^{-1}(\mathfrak{Q})$, whereas the set of priors for a single $\mathbb{Q}$ is written $\Psi^{-1}(\mathbb{Q})$.

In terms of the quantities defined above, the key quantities in [3] are

$$\begin{aligned} \mathbb{D}(f, \mu)[B] &= p_{\mathbb{D}}(B|\theta) \\ \pi \odot \mathbb{D}(A, B) &= p_{\pi, \mathbb{D}}(A, B) \\ \pi \cdot \mathbb{D}(B) &= p_{\pi, \mathbb{D}}(B). \end{aligned}$$

It may also be useful to note that $\Pi \odot_B \mathfrak{D}$ is the set of all $\pi \odot \mathbb{D}$, or equivalently, of all $p_{\pi, \mathbb{D}}(\cdot, \cdot)$, with $\pi \in \Pi$ and $\mathbb{D} \in \mathfrak{D}$, such that the marginal probability $p_{\pi, \mathbb{D}}(B)$ is greater than zero. Unless this condition holds, the posterior is undefined. $\mathcal{U}(\Pi \odot_B \mathfrak{D})$ is then our $\overline{\Phi}$, and

$\mathcal{U}(\mathcal{A})$ is the (unconditioned) supremum of $\Phi(\theta)$ for $\theta \in \mathcal{A}$, with similar definitions for $\mathcal{L}$. Here, $\mathcal{A}$ is the set of admissible parameters θ . I assume that all $\theta \in \Theta$ are admissible, so that in effect, I use Θ for $\mathcal{A}$, and ignore parameters that are not admissible.

REFERENCES

- [1] JAMES O BERGER, *An overview of robust Bayesian analysis*, TEST, 3 (1994), pp. 5–124.
- [2] T FERGUSON, *A Bayesian analysis of some nonparametric problems*, The Annals of Statistics, (1973).
- [3] HOUMAN OWHADI, CLINT SCOVEL, AND TIM SULLIVAN, *Bayesian Brittleness: Why no Bayesian model is “good enough”*, arXiv, (2013).