LA-UR-13-26482
Approved for public release; distribution is unlimited.

Title: Is Bayesian inference "brittle"?

Author(s): Wallstrom, Timothy C.
Higdon, David M.

Intended for: arXiv
Report

Issued: 2013-08-15

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity 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. Departmentof 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.

Is Bayesian inference “brittle”?[*]

Timothy C. Wallstrom[†] and David M. Higdon[‡]

Abstract. In a recent report, “Bayesian Brittleness: Why no Bayesian model is ‘good enough,’ ” (arXiv, 1304.6772v1) the authors prove rigorously that model variations that are arbitrarily small, in a particular technical sense, the posterior expectation of a function can achieve essentially any value that the function 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 explanation is that the models leading to the extreme results depend on the observed data.

1. Introduction. In order to carry out Bayesian inference, it is necessary to define the elements of the model: the prior $\pi(\theta)$ and the statistical model $p(d|\theta)$, where d is the data, and θ is a parameter. In many cases, there may be a range of modeling choices which seem more or less equally plausible. The idea of robust Bayesian inference is to consider a set of plausible priors Π, and a set of plausible models $\mathcal{M}$, and then to look at how inferences vary across these sets. If the smallest and largest inferences are similar, then the inference is “robust”: it does not depend strongly on any modeling uncertainty. 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 [1]. 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 expection 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 part of the model depends on the observed data. This part may be interpreted as either the statistical model or the prior, depending on whether we work in a parametric or nonparametric setting. If the model is allowed to depend on the data, the posterior mass can be put in any desired location, and the posterior mean of Φ can achieve any value achieved by Φ alone.

To be more specific, assume that the data take values in a continuous space, and let us

[*] 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.

[†] T-4, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, tcw@lanl.gov

[‡] CCS-6, Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, dhigdon@lanl.gov.

work in a parametric setting. The posterior mean of Φ is

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

where the posterior is given by Bayes' law,

$$\pi(\theta|d) \propto p(d|\theta)\,\pi(\theta).$$

It is easy to alter $p(\cdot|\theta)$ so that it is zero on the data for any particular value of θ, or conversely, so that it is positive for any value of θ, while keeping the original and altered versions arbitrarily close in a technical sense. Through such alterations, one can make the posterior positive on any desired subset of θ, and zero on the rest. The posterior mean can thus be engineered to have any value that can be achieved by $\Phi(\theta)$ alone.

In the nonparametric setting, θ is regarded as a measure, and $p(\cdot|\theta)$ is replaced by $\theta(\cdot)$. As before, it is easy to perturb $\theta(\cdot)$ slightly, to $\theta'(\cdot)$, so that the probability of the data is zero or positive, as desired, under the perturbed measure. Extreme results are obtained by moving the prior mass to measures θ' instead than θ, so as to achieve desired values of Φ. We do not discuss the nonparametric perspective in this paper, but a complete analysis can be found in [2].

The problem with these types of modifications is that the resulting statistical models and/or priors would not be plausible for most problems. The most implausible feature is that they are finely tuned to the exact locations of the observed data. When we consider robustness in the statistical model, we are often concerned with the possibility that, even though the shape of the distribution may be roughly correct, it may not have the exact analytic form assumed in the model. We are not seriously considering the possibility that it has notches in the exact locations of the as-yet-unobserved data. We would not even know where to put the notches before seeing the data. Even after seeing the data, we would assume that the actual locations of the data reflected statistical variability, and would not attribute these locations to the underlying model. If a statistician produced an extreme inference using such a model, it would be rejected as being based on an implausible model. Thus, these extreme inferences do not seem relevant to the question of Bayesian robustness.

In the remainder of the paper, we give two concrete examples in the parametric setting, which illustrate the basic ideas. Because we are working in the parametric setting, assumptions about the space of plausible priors, Π, correspond to assumptions about the space of possible models, $\mathcal{M}$. The analysis in [1] is carried out in a nonparametric setting, and the overall setting is much more complicated in many other ways as well. In order to isolate the essential mechanisms leading to the extreme inferences, we have suppressed many of the details, and simplified the structure of the problem somewhat. For example, the spaces $\mathcal{M}$ in the parametric setting correspond to only a subset of the space of priors Π in the nonparametric setting. A more detailed analysis, which describes the problem in a nonparametric setting and which shows more explicitly how our analysis is connected to that of Ref. [1], can be found in a longer LANL technical report [2].

2. Examples.

We illustrate the main ideas with two parametric examples.

2.1. Beta densities.

Consider a one parameter subspace of the set of beta distributions:

$$p(\cdot|\theta) = B(\cdot\,|\,c\,\theta, c(1-\theta)), \qquad \theta \in \Theta = (0,1),$$

where $c > 0$ is a constant, and

$$B(\cdot|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

Put a uniform prior density on θ, $\pi(\theta) = 1$, and compute the posterior distribution, $\pi(\theta|d)$, given n independent data points, $d = (d_1, \ldots, d_n)$. We are interested in computing the posterior mean of θ:

$$\mathbf{E}_\pi(\theta|d) = \int \theta\,\pi(\theta|d)\,d\theta,$$

which, like θ, will lie between zero and one. That is, we take $\Phi(\theta) = \theta$, which is about the simplest nontrivial possibility.

In Figure 2.1, we plot nine representative curves from the model class for the case $c = 20$, for $\theta = 0.1, 0.2, \ldots, 0.9$, together with four observed data points, $d = (0.15, 0.20, 0.28, 0.40)$. The data look as though they were generated by θ between about 0.20 and 0.40. It seems extremely unlikely that they would have come from any of the distributions with large θ, under which the probability of one or more of the points is nearly zero. Calculation gives a posterior mean of 0.26; in fact, the data was generated with $\theta = 0.30$. This calculation reflects the fact that, for this dataset, the likelihood is much higher for θ around 0.30 than for θ around 0.90, say.

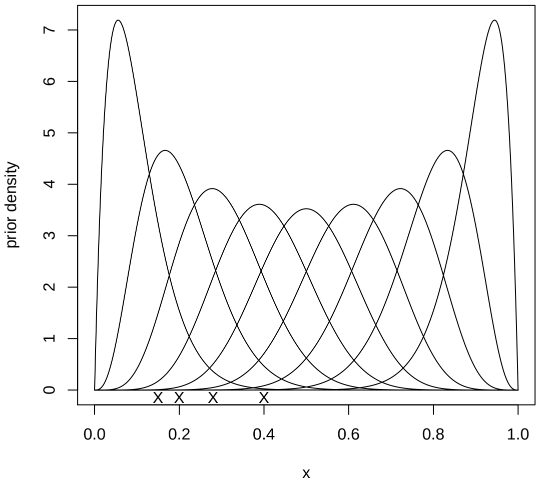

Figure 2.1. *Elements of our model class, for $\theta = 0.10, \ldots, 0.90$, together with four data points, indicated by X's along the x-axis.*

In [1], however, it is proven that when we allow even the smallest deviation from the model class, in a certain technical sense, the posterior mean can be made arbitrarily close to one, or to zero, or to any value in between, for an arbitrarily large (though finite) amount of data. Thus, it would seem that the conclusion we just reached about the posterior mean of θ is extremely fragile, and depends critically on the assumption that the true distribution is in the model class, which is never true in practice. Let us show how these extreme values are achieved for this problem, following the construction in the proof of [1, Theorem 6.10].

These extreme results are achieved, in fact, by replacing $p(\cdot|\theta)$ with $p'(\cdot|\theta)$, where, for certain values of θ, $p'(\cdot|\theta)$ has been tweaked to be zero at all of the data points; see Fig. 2.2. By Bayes' law,

$$\pi'(\theta|d) \propto \left[\prod_{i=1}^{n} p'(d_i|\theta)\right] \pi(\theta),$$

so the posterior is zero for all tweaked values of θ. Since the densities of the beta distribution are strictly positive on $(0,1)$, the posterior is positive for the remaining values of θ, and only these values contribute to the posterior mean. If we want $\mathbf{E}_\pi(\theta|d)$ to be large, we might tweak (and thereby eliminate) values of θ in $(0, 0.99)$; if we want $\mathbf{E}_\pi(\theta|d)$ to be small, we might tweak values in $(0.01, 1.00)$. The likelihood of the data may be exceedingly small for the remaining values of θ. For example, the probability of seeing four datapoints less than or equal to 0.40 for $\theta = 0.99$ is about 10^{-38}—but since the likelihood is zero elsewhere, only these values of θ will figure into the posterior mean.

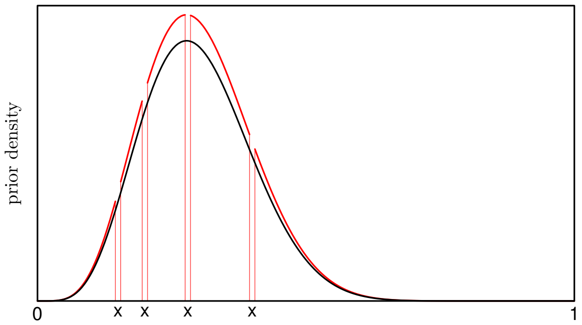

Figure 2.2. *Black:* $p(\cdot|0.30)$*; Red:* $p'(\cdot|0.30)$*, which has been set equal to zero in small neighborhoods of the data, and then renormalized to unit mass.*

By making the neighborhoods of the d_i arbitrarily small, we can ensure that the total variation distance between $p(\cdot|\theta)$ and $p'(\cdot|\theta)$ is arbitrarily small, where for continuous densities f and f',

$$d_{TV}(f, f') = \tfrac{1}{2} \int \left|f(x) - f'(x)\right| \, dx. \tag{2.1}$$

This is the technical sense in which the model perturbations are small. To be more precise, given any $\alpha > 0$, the set $\mathcal{M}$ is defined to contain, for any θ, all $p'(\cdot|\theta)$ such that

$$d_{TV}(p'(\cdot|\theta), p(\cdot|\theta)) < \alpha.$$

Although these results are mathematically correct, they seem to have little relevance to Bayesian robustness, because the tweaked models depend on the data. No one would know how to define such models before seeing the data, and in any problem we can think of, no one would consider such models realistic after seeing the data. As a result, no one would put any credence in an analysis based on such models, regardless of how close they were to the model class, in total variation distance.

2.2. Moment constraints. As a second example, let $\Theta = (0, 1)$ and $\pi(\theta) = 1$, as before, and again consider the posterior mean of θ, but let $p(\cdot|\theta)$, for each value of θ, be a distribution with mean θ. Thus, the prior predictive distribution,

$$\int p(\cdot|\theta)\,\pi(\theta)\,d\theta$$

has mean $\frac{1}{2}$. We want to show that, given an arbitrarily large number of independent data points, $d = (d_1, \ldots, d_n)$, we can choose the $p(\cdot|\theta)$ so that the posterior expectation of θ is arbitrarily close to one, regardless of the actual value of the data points. For example, even if we have a million data points, all less than 0.1, the posterior mean can be greater than 0.99.

This problem is a special case of a more general problem, in which the first k moments are constrained, where k is arbitrarily large. The intuition is that in any particular problem, we have limited information about the statistical model, and that information may be summarized, in some cases, as a knowledge of the first k moments. (We also have a prior over such statistical models, parameterized by the vector of moments.) We might imagine that if k is large and we have n datapoints, where n is large, the posterior mean of θ should be sharply defined.

We will assume, following the authors, that the true model is absolutely continuous with respect to Lebesgue measure. Also, it will be convenient, in order to avoid issues with sets of measure zero, to condition, not on the data d, but on a small ball B containing d. To be more precise, we let

$$B = B_1 \times B_2 \times \cdots \times B_n,$$

where each B_i is a small interval containing d_i. The set $\mathcal{M}$ is defined to contain, for each θ, all statistical models with mean θ.

To achieve the claimed result on the posterior mean, we follow [1, Example 5.15] in defining

$$p(\cdot|\theta) = \begin{cases} \delta_\theta(\cdot) & (\theta \le 1 - \delta) \\ (1 - \epsilon)\delta_{\theta'}(\cdot) + \epsilon\mu_d(\cdot) & (\theta > 1 - \delta), \end{cases} \tag{2.2}$$

where $\delta_\theta(\cdot)$ is the Dirac point mass at θ, μ_d is the empirical distribution of the data points,

$$\mu_d(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{d_i}(\cdot),$$

and θ' is adjusted so that the mean of $p(\cdot|\theta)$ is θ. Of course, θ' must be between zero and one, but by taking ϵ small enough, θ' can be made arbitrarily close to θ. By our continuity assumption, the d_i are distinct with probability one, and we can take the B_i to be disjoint.

We then find that for $0 < \theta \le 1 - \delta$,

$$p(B|\theta) = \prod_{i=1}^{n} p(B_i|\theta) = 0,$$

because θ cannot be in more than one disjoint set at the same time. For $\theta > 1 - \delta$, on the other hand,

$$p(B|\theta) \geq \left(\frac{\epsilon}{n}\right)^n > 0.$$

Thus, $\pi(\theta|B) > 0$ if and only if $\theta > 1 - \delta$, so that $\mathbf{E}_\pi(\theta|B) > 1 - \delta$. The example is easily extended to the case in which the first k moments are specified.

The problem, again, is that the model depends on the data, because μ_d is defined in terms of the data. In Example 2.1, the likelihood was naturally positive for any d_i, and knowledge of the data was used to choose a model such that $p(d|\theta)$ was 0 for certain values of θ. Here, the default likelihood, consisting of a Dirac point mass, was zero on any dataset with at least two distinct points, and the data was used to ensure that $p(d|\theta) > 0$ for certain values of θ.

3. Discussion. The underlying reason for the lack of robustness is that the space $\mathcal{M}$ of plausible statistical models, which might appear reasonable at an abstract mathematical level, is too large, and contains many objects that would never be adopted as statistical models in actual practice. The biggest problem is that, while the $\mathcal{M}$ are not defined in terms of the data, they are so large that they contain statistical models that are “pre-adapted” to the data, for any dataset that could possibly occur.

In the model class case, for example, $\mathcal{M}$ allows any model which is suitably close in total variation distance. However, when the data space is a continuum, any finite set of data points is generally a set of measure zero, and any particular model may be replaced by a model which is zero on all the datapoints, but which is nevertheless at TV distance zero from the original model. Thus, constraining the TV distance from the model class to be small does not ensure that the statistical models in $\mathcal{M}$ are realistic and plausible alternatives.

In the moment example, it turns out that specifying the first k moments still leaves a large space of possible statistical models. Many of these models are not very realistic, however. For example, for most problems with continuum data, a model in which all sample values are identical is implausible. But this is what is implied by the Dirac point mass models. The biggest problem, however, is that $\mathcal{M}$ contains models that depend explicitly on the data.

Acknowledgments. We thank the authors for considering these ideas, and for providing us with their feedback. We 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.

REFERENCES

[1] Houman Owhadi, Clint Scovel, and Tim Sullivan, *Bayesian Brittleness: Why no Bayesian model is“good enough”*, arXiv, (2013).

[2] Timothy C Wallstrom, *Brittleness and Bayesian Inference*, Tech. Report LA-UR-13-25883, LANL, Los Alamos, NM, July 2013.