Brittleness and Robustness in Bayesian Inference

Tim Sullivan
with Houman Owhadi and Clint Scovel (Caltech)

Mathematics Institute
University of Warwick

SIAM Conference on Uncertainty Quantification
Savannah, Georgia, U.S.A.
3 April 2014
Bayesian procedures give posterior estimates for quantities of interest in the form of Bayes’ rule

\[ p(\text{parameters}|\text{data}) \propto L(\text{data}|\text{parameters})p(\text{parameters}) \]

given the following data:
- a prior probability distribution on parameters — later denoted \( x \in X \);
- a likelihood function;
- observations/data/evidence — later denoted \( y \in Y \).

It is natural to ask about the robustness, stability, and accuracy of such procedures.

This is a subtle topic, with both positive and negative results, especially for large/complex systems, with fine geometrical and topological considerations playing a key role.
What Do We Mean by ‘Bayesian Brittleness’?

\[ p(\text{parameters}|\text{data}) \propto L(\text{data}|\text{parameters})p(\text{parameters}) \]

- **Frequentist** questions: If the data are generated from some ‘true’ distribution, will the posterior eventually/asymptotically identify the ‘true’ value? Are Bayesian credible sets also frequentist confidence sets? What if the model class doesn’t even contain the ‘truth’?

- **Numerical analysis** questions: Is Bayesian inference a well-posed problem, in the sense that small perturbations of the prior, likelihood, or data (e.g. those arising from numerical discretization) lead to small changes in the posterior? Can effective estimates be given?

- For us, ‘brittleness’ simply means the **strongest possible negative result**: under arbitrarily small perturbations of the problem setup the posterior conclusions change as much as possible — i.e. extreme discontinuity. (More precise definition later on.)
Overview

1. Bayesian-Frequentist Consistency

2. Bayesian Brittleness

3. Closing Remarks
Overview

1. Bayesian-Frequentist Consistency
2. Bayesian Brittleness
3. Closing Remarks
Parameter space $\mathbb{X}$, equipped with a prior $\mu_0 \in \mathcal{P}(\mathbb{X})$.

We observe data with values in $\mathbb{Y}$.

The Bayesian model (or likelihood) is a function $L: \mathbb{X} \to \mathcal{P}(\mathbb{Y})$, i.e.

$$L(E|x) = \mathbb{P}[y \in E \mid x].$$

This defines a (non-product) measure $\mu$ on $\mathbb{X} \times \mathbb{Y}$ by

$$\mu(E) := \mathbb{E}_{x \sim \mu_0, y \sim L(\cdot \mid x)}[1_E(x, y)] \equiv \int_{\mathbb{X}} \int_{\mathbb{Y}} 1_E(x, y) L(dy|x) \mu_0(dx).$$

The Bayesian posterior on $\mathbb{X}$ is just $\mu$ conditioned on a $\mathbb{Y}$-fibre.
Parameter space $\mathbb{X}$, equipped with a prior $\mu_0 \in \mathcal{P}(\mathbb{X})$.

We observe data with values in $\mathbb{Y}$.

The Bayesian model (or likelihood) is a function $L: \mathbb{X} \to \mathcal{P}(\mathbb{Y})$, i.e.

$$L(E|x) = \mathbb{P}[y \in E | x].$$

**Definition (Frequentist well-specification)**

If data are generated according to $\mu^\dagger \in \mathcal{P}(\mathbb{Y})$, then the Bayesian model is called **well-specified** if there is some $x^\dagger \in \mathbb{X}$ such that $\mu^\dagger = L(\cdot | x^\dagger)$; otherwise, the model is called **misspecified**.
Consistency of Bayesian Models

Suppose that the observed data consists of a sequence of independent \( \mu^\dagger \)-distributed samples \((y_1, y_2, \ldots)\), and let

\[
\mu_n(x) := \mu_0(x | y_1, \ldots, y_n) \propto L(y_1, \ldots, y_n | x) \mu_0(x)
\]

be the posterior measure obtained by conditioning the prior \(\mu_0\) with respect to the first \(n\) observations using Bayes’ rule.

**Definition (Frequentist consistency)**

A well-specified Bayesian model with \(\mu^\dagger = L(\cdot | x^\dagger)\) is called **consistent** (in an appropriate topology on \(\mathcal{P}(X)\)) if

\[
\lim_{n \to \infty} \mu_n = \delta_{x^\dagger},
\]

i.e. the posterior asymptotically gives full mass to the true parameter value.
The classic positive result regarding posterior consistency is the Bernstein–von Mises theorem or Bayesian CLT, historically first envisioned by Laplace (1810) and first rigorously proved by Le Cam (1953):

**Theorem (Bernstein–von Mises)**

If \( \mathbb{X} \) and \( \mathbb{Y} \) are finite-dimensional, then, subject to regularity assumptions on \( L \), any well-specified Bayesian model is consistent, so long as \( x^\dagger \in \text{supp}(\mu_0) \). Furthermore, \( \mu_n \) is asymptotically normal with precision proportional to the Fisher information matrix \( \mathcal{I}(x^\dagger) \):

\[
\mathbb{P}_{Y_i \sim \mu^\dagger} \left[ \left\| - \frac{1}{n} \mathcal{I}(x^\dagger)^{-1} \hat{x}_n^{\text{MLE}} - \mathcal{N}(\mu_n, \mathcal{I}(x^\dagger)^{-1}) \right\|_{TV} > \varepsilon \right] \longrightarrow 0, \quad n \to \infty,
\]

where

\[
\mathcal{I}(x^\dagger)_{ij} = \mathbb{E}_{y \sim L(\cdot|x^\dagger)} \left[ \frac{\partial \log L(y|x)}{\partial x_i} \frac{\partial \log L(y|x)}{\partial x_j} \bigg|_{x=x^\dagger} \right].
\]
Informally, the BvM theorem says that a well-specified model is capable of learning any ‘truth’ in the support of the prior.

If we obey Cromwell’s Rule

“I beseech you, in the bowels of Christ, think it possible that you may be mistaken.”

by choosing a globally supported prior $\mu_0$, then everything should turn out OK — and the limiting posterior should be independent of $\mu_0$.

Unfortunately, the BvM theorem is not always true if $\dim \mathcal{X} = \infty$, even for globally supported priors — but nor is it always false.

Applications of Bayesian methods in function spaces are increasingly popular, so it is important to understand the precise circumstances in which we do or do not have the BvM property.
## Positive and Negative Consistency Results

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical BvM</td>
<td><strong>Freedman</strong> (1963, 1965)</td>
</tr>
<tr>
<td><strong>Castillo &amp; Rousseau</strong> (2013) and <strong>Nickl &amp; Castillo</strong> (2013): Gaussian seq. space model, modified $\ell^2$ balls</td>
<td><strong>Diaconis &amp; Freedman</strong> (1998)</td>
</tr>
<tr>
<td><strong>Stuart &amp; al.</strong> (2010+): Gaussian and Besov measures on Banach spaces, Hellinger metric</td>
<td><strong>Johnstone</strong> (2010) and <strong>Leahu</strong> (2011): further Freedman-type results</td>
</tr>
<tr>
<td></td>
<td>Owhadi, Scovel &amp; S.</td>
</tr>
</tbody>
</table>

The main moral to be drawn from this zoo of theorems and examples is that the geometry and topology of the spaces play a critical role in consistency properties.
By definition, if the model is mis-specified, then we cannot hope for posterior consistency in the sense that $\mu_n \to \delta_{x^\dagger}$ where $L(\cdot | x^\dagger) = \mu^\dagger$, because no such $x^\dagger \in \mathcal{X}$ exists.

However, we can still hope that $\mu_n \to \delta_{\hat{x}}$ for some ‘meaningful’ $\hat{x} \in \mathcal{X}$, and that we get consistent estimates for the values of suitable quantities of interest, e.g. the posterior asymptotically puts all mass on $\hat{x} \in \mathcal{X}$ such that $L(\cdot | \hat{x})$ matches the mean and variance of $\mu^\dagger$, if not the exact distribution.

For example, Berk (1966, 1970), Kleijn & Van der Vaart (2006), Shalizi (2009) have results of the type:

**Theorem (Minimum relative entropy)**

*Under suitable regularity assumptions, the posterior concentrates on*

$$\hat{x} \in \arg \min_{x \in \text{supp}(\mu_0)} D_{KL}(\mu^\dagger \| L(\cdot | x)).$$
Overview

1. Bayesian-Frequentist Consistency

2. Bayesian Brittleness

3. Closing Remarks
For simplicity, $\mathbb{X}$ and $\mathbb{Y}$ will be separable metric spaces — see arXiv:1304.6772 for weaker but more verbose assumptions.

Fix a prior $\mu_0 \in \mathcal{P}(\mathbb{X})$; we will consider other priors $\tilde{\mu}_0$ ‘near’ to $\mu_0$.

Consider a *misspecified* likelihood model.

Given $\mu_0$ and any quantity of interest $q : \mathbb{X} \to \mathbb{R}$,

$$
\mu_0\text{-ess inf}_{x \in \mathbb{X}} q(x) := \sup \left\{ t \in \mathbb{R} \mid q(x) \geq t \ \mu_0\text{-a.s.} \right\},
$$

$$
\mu_0\text{-ess sup}_{x \in \mathbb{X}} q(x) := \inf \left\{ t \in \mathbb{R} \mid q(x) \leq t \ \mu_0\text{-a.s.} \right\}.
$$

To get around difficulties of data actually having measure zero, and with one eye on the fact that real-world data is always discretized to some precision level $0 < \delta < \infty$, we assume that our observation is actually that the ‘exact’ data lies in a metric ball $B_\delta(y) \subseteq \mathbb{Y}$.

Slight modification: $y$ could actually be $(y_1, \ldots, y_n) \in \mathbb{Y}^n$. 
Setup for Brittleness

The brittleness theorem covers three notions of closeness:

- **total variation distance**: for $\alpha > 0$ (small), $\|\mu_0 - \tilde{\mu}_0\|_{TV} < \alpha$; or
- **Prohorov distance**: for $\alpha > 0$ (small), $d_\Pi(\mu_0, \tilde{\mu}_0) < \alpha$ (for separable $\mathbb{X}$, this metrizes the weak convergence topology on $\mathcal{P}(\mathbb{X})$); or
- **common moments**: for $\alpha \in \mathbb{N}$ (large), for prescribed measurable functions $\phi_1, \ldots, \phi_\alpha : \mathbb{X} \to \mathbb{R}$,

  $$
  E_{\mu_0}[\phi_i] = E_{\tilde{\mu}_0}[\phi_i] \quad \text{for } i = 1, \ldots, \alpha,
  $$

  or, for $\varepsilon_i > 0$,

  $$
  |E_{\mu_0}[\phi_i] - E_{\tilde{\mu}_0}[\phi_i]| \leq \varepsilon_i \quad \text{for } i = 1, \ldots, \alpha.
  $$
Theorem (Brittleness)

Suppose that a misspecified model permits observed data to be arbitrarily unlikely in the sense that

\[
\lim_{\delta \to 0} \sup_{y \in Y} \sup_{x \in X} L(B_\delta(y) \mid x) = 0, \tag{AU}
\]

and let \( q: \mathbb{X} \to \mathbb{R} \) be any measurable function. Then, for all

\[
v \in \left[ \mu_0\text{-ess inf}_{x \in X} q(x), \mu_0\text{-ess sup}_{x \in X} q(x) \right],
\]

and all \( \alpha > 0 \), there exists \( \delta_*(\alpha) > 0 \) and a prior \( \tilde{\mu}_0 \) ‘\( \alpha \)-close’ to \( \mu_0 \) such that the posterior value \( \mathbb{E}_{\tilde{\mu}_0} \left[ q \mid B_\delta(y) \right] \) for \( q \) given data of precision \( 0 < \delta < \delta_*(\alpha) \) is the chosen value \( v \).
Idea of Proof

- Optimize over the set $A_\alpha$ of priors that are $\alpha$-close to the original prior. (*Cf.* construction of Bayesian least favourable priors and frequentist minimax estimators.)

- The three notions of closeness considered (moments, Prokhorov, TV), plus the misspecification assumption, plus the (AU) condition, together permit priors $\tilde{\mu}_0 \in A_\alpha$ to put lots of posterior mass on ‘bad’ data sets.

- In our proof as written, the perturbations used to produce the ‘bad’ priors use point masses; a slight variation would produce the same result using absolutely continuous perturbations.
Schematically, the perturbation from $\mu_0$ to $\tilde{\mu}_0$ looks like

What would break this argument? Stronger topologies, e.g. closeness in the Kullback–Leibler or Hellinger distances. However, finiteness of these distances requires mutual absolute continuity of the two measures considered — which is a lot to ask for in infinite-dimensional cases.
Schematically, the perturbation from $\mu_0$ to $\tilde{\mu}_0$ looks like

What would break this argument? Stronger topologies, e.g. closeness in the Kullback–Leibler or Hellinger distances. However, finiteness of these distances requires mutual absolute continuity of the two measures considered — which is a lot to ask for in infinite-dimensional cases.
Misspecification has profound consequences for Bayesian robustness on ‘large’ spaces — in fact, Bayesian inferences become extremely brittle as a function of measurement resolution $\delta$.

If the model is misspecified, and there are possible observed data that are arbitrarily unlikely under the model, then under fine enough measurement resolution the posterior predictions of nearby priors differ as much as possible regardless of the number of samples observed.

Figure. As measurement resolution $\delta \to 0$, the smooth dependence of $\mathbb{E}_{\mu_0}[q]$ on the prior $\mu_0$ (top-left) shatters into a patchwork of diametrically opposed posterior values $\mathbb{E}_{\mu_n}[q] \equiv \mathbb{E}_{\mu_0}[q|y_1, \ldots, y_n]$. 
Overview

1. Bayesian-Frequentist Consistency

2. Bayesian Brittleness

3. Closing Remarks
In contrast to the classical robustness and consistency results for Bayesian inference for discrete or finite-dimensional systems, the situation for infinite-dimensional spaces is *complicated*.

Bayesian inference is *extremely brittle in some topologies*, and so cannot be consistent, and high-precision data only worsens things.

**Consistency can hold for complex systems**, with *careful* choices of prior, geometry and topology — but, since the situation is so sensitive, *all assumptions must be considered carefully*.

*In practice, Bayesian inference is employed under misspecification *all the time*, particularly so in machine learning applications. While sometimes it works quite well under misspecification, there are also cases where it does not, so it seems important to determine precise conditions under which misspecification is harmful — even if such an analysis is based on frequentist assumptions.*

— P. D. Grünwald. “Bayesian Inconsistency under Misspecification.”
Thank You