Brittleness and Robustness of Bayesian Inference

Tim Sullivan[1]
with Houman Owhadi[2] and Clint Scovel[2]

[1]Free University of Berlin / Zuse Institute Berlin, Germany
[2]California Institute of Technology, USA

Joint Analysis Seminar
RWTH Aachen University, Germany
24 November 2015

What Do We Mean by 'Bayesian Brittleness'?

- Bayesian procedures give posterior distributions 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 $u \in \mathbb{U}$;
    - a likelihood function;
    - observations / data — later denoted $y \in \mathbb{Y}$.
- 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

Bayesian Modelling Setup

- Parameter space $\mathbb{U}$, equipped with a prior $\pi \in \mathcal{P}(\mathbb{U})$.
- Observed data with values in $\mathbb{Y}$ are explained using a likelihood model, i.e. a function $L \colon \mathbb{U} \to \mathcal{P}(\mathbb{Y})$ with

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

- This defines a (non-product) joint measure μ on $\mathbb{U} \times \mathbb{Y}$ by

$$\mu(E) := \mathbb{E}_{u \sim \pi, y \sim L(\,\cdot\,|u)}\big[\mathbb{1}_E(u, y) \big] \equiv \int_{\mathbb{U}} \int_{\mathbb{Y}} \mathbb{1}_E(u, y)\, L(\mathrm{d}y|u)\, \pi(\mathrm{d}u).$$

- The Bayesian posterior on $\mathbb{U}$ is just μ conditioned on a $\mathbb{Y}$-fibre, and re-normalized to be a probability measure. Bayes' Rule gives this as

$$\mathbb{P}(u|y) = \frac{\mathbb{P}[y|u]\mathbb{P}[u]}{\mathbb{P}[y]}.$$

Bayesian Modelling Setup

- Parameter space $\mathbb{U}$, equipped with a prior $\pi \in \mathcal{P}(\mathbb{U})$.
- Observed data with values in $\mathbb{Y}$ are explained using a likelihood model, i.e. a function $L \colon \mathbb{U} \to \mathcal{P}(\mathbb{Y})$ with

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

- This defines a (non-product) joint measure μ on $\mathbb{U} \times \mathbb{Y}$ by

$$\mu(E) := \mathbb{E}_{u \sim \pi, y \sim L(\,\cdot\,|u)}\big[\mathbb{1}_E(u, y)\big] \equiv \int_{\mathbb{U}} \int_{\mathbb{Y}} \mathbb{1}_E(u, y)\, L(\mathrm{d}y|u)\, \pi(\mathrm{d}u).$$

- The Bayesian posterior on $\mathbb{U}$ is just μ conditioned on a $\mathbb{Y}$-fibre, and re-normalized to be a probability measure. Bayes' Rule gives this as

$$\pi(E|y) = \frac{\mathbb{E}_{u \sim \pi}[\mathbb{1}_E(u) L(y|u)]}{\mathbb{E}_{u \sim \pi}[L(y|u)]}.$$

Bayesian Modelling Setup

- Parameter space $\mathbb{U}$, equipped with a prior $\pi \in \mathcal{P}(\mathbb{U})$.
- Observed data with values in $\mathbb{Y}$ are explained using a likelihood model, i.e. a function $L \colon \mathbb{U} \to \mathcal{P}(\mathbb{Y})$ with

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

- This defines a (non-product) joint measure μ on $\mathbb{U} \times \mathbb{Y}$ by

$$\mu(E) := \mathbb{E}_{u \sim \pi, y \sim L(\,\cdot\,|u)}\big[\mathbb{1}_E(u, y)\big] \equiv \int_{\mathbb{U}} \int_{\mathbb{Y}} \mathbb{1}_E(u, y)\, L(\mathrm{d}y|u)\, \pi(\mathrm{d}u).$$

- The Bayesian posterior on $\mathbb{U}$ is just μ conditioned on a $\mathbb{Y}$-fibre, and re-normalized to be a probability measure. Bayes' Rule gives this as

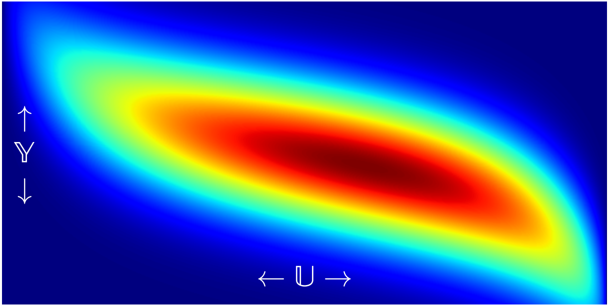
$$\frac{\mathrm{d}\pi(\,\cdot\,|y)}{\mathrm{d}\pi} \propto L(y|\,\cdot\,).$$

Bayesian Modelling Setup

Prior measure π on $\mathbb{U}$:

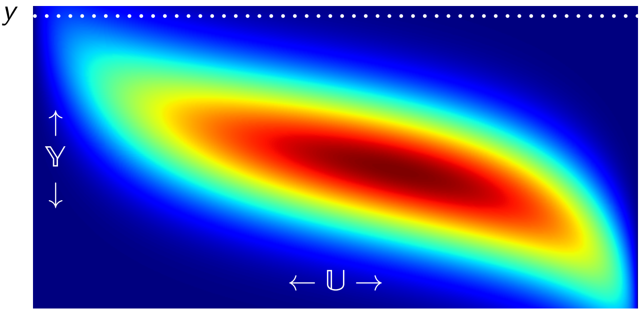
Prior measure π on $\mathbb{U}$:

Joint measure μ on $\mathbb{U} \times \mathbb{Y}$:



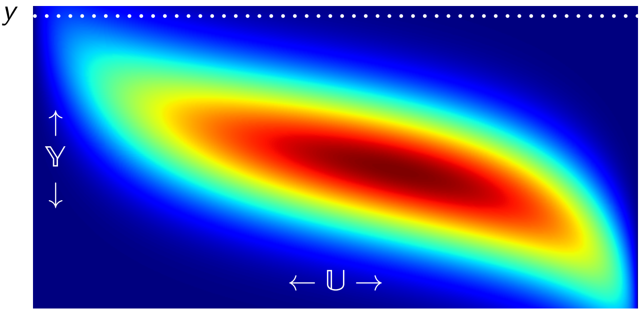
Prior measure π on $\mathbb{U}$:

Joint measure μ on $\mathbb{U} \times \mathbb{Y}$:



Prior measure π on $\mathbb{U}$:

Joint measure μ on $\mathbb{U} \times \mathbb{Y}$:



Posterior measure $\pi(\cdot|y) \propto \mu|_{\mathbb{U}\times\{y\}}$ on $\mathbb{U}$:

Bayesian Modelling Setup

Traditional Setting

$\mathbb{U}$ is a finite set or $\mathbb{R}^d$ for small $d \in \mathbb{N}$.

Finite-Dimensional Linear / Gaussian Prototype

- Linear observation equation with additive noise η:

$$y = Hu + \eta.$$

- This defines the conditional distribution of $y|u$, to be inverted using Bayes' rule to obtain the posterior $u|y$.
- If the prior on u is Gaussian, $H \colon \mathbb{U} \to \mathbb{Y}$ is linear, and $\eta \sim \mathcal{N}(m, Q)$ is Gaussian, then (u, y), $y|u$, and $u|y$ are all Gaussian. $(L(y|u) \propto \exp(-\frac{1}{2}\|y - Hu\|_Q^2))$
- Get the posterior mean and covariance by Schur complementation / Woodbury lemma.

Bayesian Modelling Setup

Traditional Setting

$\mathbb{U}$ is a finite set or $\mathbb{R}^d$ for small $d \in \mathbb{N}$.

More Modern Applications

A very high-dimensional or infinite-dimensional $\mathbb{U}$, e.g. an inverse problem for a PDE:

$$\begin{aligned} -\nabla \cdot (u \nabla p) &= f, \\ \text{boundary conditions}(p) &= 0. \end{aligned}$$

in which we attempt to infer the permeability u from e.g. some noisy point observations of the pressure/head p. Priors are Gaussian, Besov, or other measures on function spaces; posteriors are typically complicated.

Specification of Bayesian Models

- Parameter space $\mathbb{U}$, equipped with a prior $\pi \in \mathcal{P}(\mathbb{U})$.
- Observed data with values in $\mathbb{Y}$ are explained using a likelihood model, i.e. a function $L\colon \mathbb{U} \to \mathcal{P}(\mathbb{Y})$ with

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

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 $u^\dagger \in \mathbb{U}$ such that $\mu^\dagger = L(\,\cdot\,|u^\dagger)$; 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 $(y_1, y_2, \dots)$, and let

$$\pi^{(n)}(u) := \pi(u|y_1, \dots, y_n) \propto L(y_1, \dots, y_n|u)\pi(u)$$

be the posterior measure obtained by conditioning the prior π with respect to the first n observations using Bayes' rule.

Definition (Frequentist consistency)

A well-specified Bayesian model with $\mu^\dagger = L(\,\cdot\,|u^\dagger)$ is called consistent (in an appropriate topology on $\mathcal{P}(\mathbb{U})$) if

$$\lim_{n\to\infty} \pi^{(n)} = \delta_{u^\dagger},$$

i.e. the posterior asymptotically gives full mass to the true parameter value.

Bernstein–von Mises Theorem

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 $\mathbb{U}$ and $\mathbb{Y}$ are finite-dimensional, then, subject to regularity assumptions on L and π, any well-specified Bayesian model is consistent provided $u^\dagger \in \operatorname{supp}(\pi)$. Furthermore, $\pi^{(n)}$ is asymptotically normal about $\widehat{u}_n^{\mathrm{MLE}} \to u^\dagger$, with precision proportional to the Fisher information $\mathcal{I}(u^\dagger)$:

$$\mathbb{P}_{y_i \sim \mu^\dagger} \left[\left\| \pi^{(n)} - \mathcal{N}\left(\widehat{u}_n^{\mathrm{MLE}}, \frac{\mathcal{I}(u^\dagger)^{-1}}{n} \right) \right\|_{\mathrm{TV}} > \varepsilon \right] \xrightarrow[n \to \infty]{} 0,$$

where $$\mathcal{I}(u^\dagger)_{ij} = \mathbb{E}_{y \sim L(\,\cdot\,|u^\dagger)} \left[\frac{\partial \log L(y|u)}{\partial u_i} \frac{\partial \log L(y|u)}{\partial u_j} \bigg|_{u=u^\dagger} \right].$$

Bernstein–von Mises Theorem

- 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 (**Lindley**, 1965)

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

  by choosing a globally supported prior π, then everything should turn out OK — and the limiting posterior should be independent of π.
- Unfortunately, the BvM theorem is not always true if $\dim \mathbb{U} = \infty$, 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.

Some Positive and Negative Consistency Results

Positive

- **Barron, Schervish & Wasserman** (1999): K–L and Hellinger
- **Castillo & Rousseau** and **Nickl & Castillo** (2013): Gaussian seq. space model, modified ℓ^2 balls
- **Szabó, Van der Vaart, Van Zanten** (2014)
- **Stuart** & al. (2010+): Gaussian / Besov measures
- Dirichlet processes

Negative

- **Freedman** (1963, 1965): prior supported on $\mathcal{P}(\mathbb{N}_0)$ sees i.i.d. $y_i \sim \mathrm{Geom}(\frac{1}{4})$, but posterior $\to \mathrm{Geom}(\frac{3}{4})$
- **Diaconis & Freedman** (1998): such 'bad' priors are of small measure, but are topologically generic
- **Johnstone** (2010) and **Leahu** (2011): further Freedman-type results
- $\to$ **Owhadi, Scovel & S.**

Main moral: the geometry and topology play a critical role in consistency.

- By definition, if the model is mis-specified, then we cannot hope for posterior consistency in the sense that $\pi^{(n)} \to \delta_{u^\dagger}$ where $L(\cdot|u^\dagger) = \mu^\dagger$, because no such $u^\dagger \in \mathbb{U}$ exists.
- However, we can still hope that $\pi^{(n)} \to \delta_{\widehat{u}}$ for some 'meaningful' $\widehat{u} \in \mathbb{U}$, and that we get consistent estimates for the values of suitable quantities of interest, e.g. the posterior asymptotically puts all mass on $\widehat{u} \in \mathbb{U}$ such that $L(\cdot|\widehat{u})$ 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

$$\widehat{u} \in \operatorname{arg\,min} \left\{ D_{\mathrm{KL}}\left(\mu^\dagger \middle\| L(\cdot|u)\right) \middle| u \in \operatorname{supp}(\pi_0) \right\}.$$

Overview

Setup for Brittleness

The brittleness theorem covers three notions of closeness between models:

- total variation distance: for $\lambda > 0$ (small), $\|\mu_0 - \mu_\lambda\|_{\mathsf{TV}} < \lambda$; or
- Prohorov distance: for $\lambda > 0$ (small), $d_\Pi(\mu_0, \mu_\lambda) < \lambda$ (for separable $\mathbb{U}$, this metrizes the weak convergence topology on $\mathcal{P}(\mathbb{U})$); or
- common moments: for $\lambda \in \mathbb{N}$ (large), for prescribed measurable functions $\phi_1, \dots, \phi_\lambda \colon \mathbb{U} \times \mathbb{Y} \to \mathbb{R}$,

$$\mathbb{E}_{\mu_0}[\phi_i] = \mathbb{E}_{\mu_\lambda}[\phi_i] \quad \text{for } i = 1, \dots, \lambda,$$

or, for $\varepsilon_i > 0$,

$$\bigl|\mathbb{E}_{\mu_0}[\phi_i] - \mathbb{E}_{\mu_\lambda}[\phi_i]\bigr| \leq \varepsilon_i \quad \text{for } i = 1, \dots, \lambda.$$

Setup for Brittleness

- For simplicity, $\mathbb{U}$ and $\mathbb{Y}$ will be complete and separable metric spaces — see arXiv:1304.6772 for weaker but more verbose assumptions.
- Fix a prior $\pi_0 \in \mathcal{P}(\mathbb{U})$ and likehood model L_0, and the induced joint measure (Bayesian model) μ_0; we will consider other models μ_λ 'near' to μ_0.
- Given π_0 and any quantity of interest $q \colon \mathbb{U} \to \mathbb{R}$,

$$\pi_0\text{-}\operatorname*{ess\,inf}_{u \in \mathbb{U}} q(u) := \sup\left\{ t \in \mathbb{R} \,\middle|\, q(u) \geq t \ \pi_0\text{-a.s.} \right\},$$
$$\pi_0\text{-}\operatorname*{ess\,sup}_{u \in \mathbb{U}} q(u) := \inf\left\{ t \in \mathbb{R} \,\middle|\, q(u) \leq t \ \pi_0\text{-a.s.} \right\}.$$

- 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 $\mathbb{B}_\delta(y) \subseteq \mathbb{Y}$.
- Slight modification: y could actually be $(y_1, \dots, y_n) \in \mathbb{Y}^n$.

Brittleness Theorem

Owhadi, Scovel & S. *Elec. J. Stat.*, 2015, Theorem 6.4

Theorem (Brittleness)

Suppose that the original model (π_0, L_0) *is mis-specified and permits observed data to be arbitrarily unlikely in the sense that*

$$\lim_{\delta \to 0} \sup_{\substack{y \in \mathbb{Y} \\ u \in \operatorname{supp}(\pi_0) \subseteq \mathbb{U}}} L_0\big(\mathbb{B}_\delta(y) \,\big|\, u \big) = 0, \tag{AU}$$

and let $q \colon \mathbb{U} \to \mathbb{R}$ *be any measurable function. Then, for all*

$$v \in \left[\pi_0\text{-}\operatorname*{ess\,inf}_{u \in \mathbb{U}} q(u), \pi_0\text{-}\operatorname*{ess\,sup}_{u \in \mathbb{U}} q(u) \right],$$

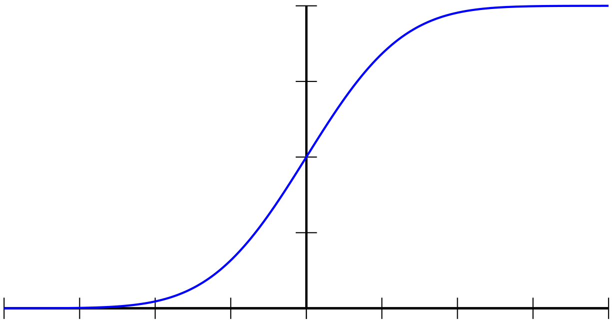
and all $\lambda > 0$*, there exists* $\delta_*(\lambda) > 0$ *and* μ_λ *'*λ*-close' to* μ_0 *such that, for all* $0 < \delta < \delta_*(\lambda)$ *and all* $y = (y_1, \dots, y_n) \in \mathbb{Y}^n$*, the posterior value* $\mathbb{E}_{\pi_\lambda}\big[q \,\big|\, \mathbb{B}_\delta(y) \big]$ *is the chosen value* v.

Idea of Proof

- Optimize over the set $\mathcal{A}_\lambda$ of models that are λ-close to the original model. (*Cf.* construction of Bayesian least favourable priors and frequentist minimax estimators.)
- This involves understanding extreme points of $\mathcal{A}_\lambda$ and the optimization of affine functionals over such sets — Choquet theory and results of **von Weizsäcker & Winkler** — and previous work by S. and collaborators on Optimal UQ (*SIAM Rev.*, 2013).
- The three notions of closeness considered (moments, Prohorov, TV), plus the (AU) condition, together permit models $\mu_\lambda \in \mathcal{A}_\lambda$ to 'choose which data to trust' when forming the posterior.
- In our proof as written, the perturbations used to produce the 'bad' models use point masses; a slight variation would produce the same result using absolutely continuous perturbations.

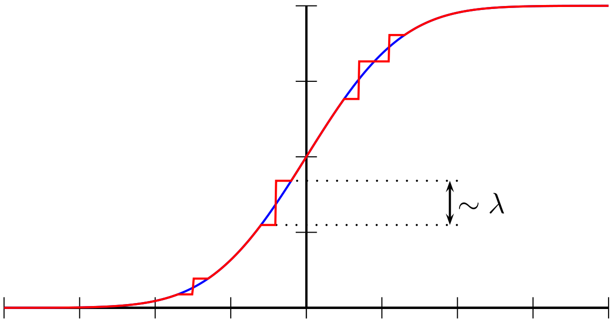
Brittleness Theorem — Proof Sketch

Schematically, the perturbation from μ_0 to μ_λ looks like



Brittleness Theorem — Proof Sketch

Schematically, the perturbation from μ_0 to μ_λ looks like



Brittleness Theorem — Interpretation

- Misspecification has strong consequences for Bayesian robustness on 'large' spaces — in fact, Bayesian inferences become extremely brittle as a function of measurement resolution δ.
- 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*.

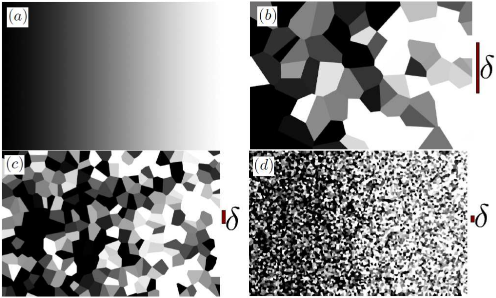


Figure. As measurement resolution $\delta \to 0$, the smooth dependence of $\mathbb{E}_{\pi_0}[q]$ on the prior π_0 (top-left) shatters into a patchwork of diametrically opposed posterior values $\mathbb{E}_{\pi^{(n)}}[q] \equiv \mathbb{E}_{\pi_0}[q|\mathbb{B}_\delta(y)]$.

Brittleness Rates — A Moment-Based Example

- Estimate the mean of a random variable X, taking values in $[0, 1]$, given a single observation y of X
- Set $\mathcal{A}$ of admissible priors for the law of X: anything that gives uniform measure to the mean, uniform measure to the second moment given the mean, uniform measure to the third moment given the second, ... up to k^{th} moment. (Note that $\dim \mathcal{A} = \infty$ but $\operatorname{codim} \mathcal{A} = k$.)
- So, in particular, for any prior $\pi \in \mathcal{A}$, $\mathbb{E}_\pi[\mathbb{E}[X]] = \frac{1}{2}$.
- Can find priors $\pi_1, \pi_2 \in \mathcal{A}$ with

$$\mathbb{E}_{\pi_1}[\mathbb{E}[X]|y] \leq 4e \left[\frac{2k\delta}{e} \right]^{\frac{1}{2k+1}} \approx 0,$$

$$\mathbb{E}_{\pi_2}[\mathbb{E}[X]|y] \geq 1 - 4e \left[\frac{2k\delta}{e} \right]^{\frac{1}{2k+1}} \approx 1.$$

- Taking δ fixed and $n \to \infty$ does not prevent brittleness in the classical Bayesian sensitivity analysis framework (it only leads to more directions of instabilities).
- For a fixed δ, the previous example suggests that brittleness results do not persist in that same framework when the number of moment constraints k (on the class of priors) is large enough.
- Furthermore, taking $\delta > 0$ fixed (or discretizing space at a resolution $\delta > 0$) enables the construction of classes of qualitatively robust priors (to TV perturbations) that are nearly consistent as $n \to \infty$ (some degree of consistency is lost due to the dis- cretization).
- Restricting to finite-precision data, i.e. keeping δ bounded away from zero, is physically quite reasonable. The universe may be granular enough that $\delta^{1/(2k+1)} \gg 0$ for all 'practical' $\delta > 0$.

Overview

The Answer Will Not be Purely Bayesian

- A close inspection of some of the cases where Bayesian inference has been successful suggests the existence of a non-Bayesian feedback loop on the evaluation of its performance, e.g. frequentist methods, posterior predictive checking. (**Gelman & al.**, 1996; **Mayo & Spanos**, 2004)

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."

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.
- And, once a 'mathematical' prior is agreed upon, just as with classical numerical analysis of algorithms for ODEs and PDEs, the onus is on the algorithm designer to ensure that 'numerical' prior is close to the 'mathematical' one in a 'good' topology.

Thank You

- H. Owhadi & C. Scovel, arXiv:1304.7046
- H. Owhadi, C. Scovel & T. J. Sullivan, *SIAM Rev.* **57**(4):566–582, 2015.
- H. Owhadi, C. Scovel & T. J. Sullivan, *Elec. J. Stat.* **9**:1–79, 2015.
- H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns & M. Ortiz, *SIAM Rev.* **55**(2):271–345, 2013.
- T. J. Sullivan, M. McKerns, D. Meyer, F. Theil, H. Owhadi & M. Ortiz, *Math. Model. Numer. Anal.* **47**(6):1657–1689, 2013.