Robust Bayesian Updating

Jack Jewson (Warwick), Jim Smith (Warwick) and Chris Holmes (Oxford)

Objectives

1. Demonstrate lack of robustness in traditional Bayesian inference.
2. Introduce Bayesian updating using a loss function.
3. Propose a more robust Bayesian update to be used in the M-open world.

Background

The Bayesian decision problem:

- choose a decision $d \in \mathcal{D}$,
- in order to minimise $\ell(d, x)$,
- against some future unknown $x \in \mathcal{X}$,

Bayesian statistics provides the tools to solve.

- Savage:
  - probabilities are beliefs,
  - preferences are loss functions,
  - the optimal decision minimises the expected loss.
- Inference is a decision problem where the decision is a probability distribution [1].
- The log-score is local and proper.
- Kullback-Leibler (KL) divergence $d_{KL}(g||f)$, is the expected log-score of deciding f when g is the truth.

The Problem

"If preferences are described by the log score, one should beware of approximating by 0" [1]

- As $x \to 0$, $-\log(x) \to \infty$.
- Severe penalty for predicting an observed event to have probability close to 0.
- Results in a desire to correctly capture the tail behaviour of the data generating process.
- Important for pure inference problems [1].
- In the M-open world the model is never correct.
- Leads to Bayesian updating being very non-robust.
- e.g. under ϵ-contamination the KL-divergence is unbounded. (see 'Demonstration')

General Bayesian Updating

- Decision problem (parametrised by θ)
- The 'true' Bayes act:

$$\theta^* = \arg\min_{\theta} \int_{\mathcal{X}} \ell(\theta, x) dG, \tag{1}$$

  where $G(x)$ is the true data generating density.
- The traditional Bayesian builds a belief model to approximate $G(x)$.

Without a model, the General Bayesian [2] posterior must be close to:

- the prior (measured using KL-divergence)
- and the data (measured using expected loss)

The posterior minimising the sum of these is:

$$\pi(\theta|\mathbf{x}) \propto \exp\{-w \sum_i \ell(\theta, x_i)\}\pi(\theta) \tag{2}$$

- High posterior mass is assigned to parameters minimising the loss given the data.
- The data is used to empirically integrate over $G(x)$.

Bayes as General Bayes

If $\ell(\theta, x) = -\log(f(x; \theta))$ then the general Bayesian update recovers Bayes rule, in agreement with [1]:

$$\pi(\theta|\mathbf{x}) \propto \prod_i \{f(x_i; \theta)\}\pi(\theta) \tag{3}$$

Hellinger Bayes Issues

- No longer using Bayes rule, so need to correctly set w to ensure the H-Bayes posterior maintains probabilistic meaning.
- No longer have the likelihood principle or Bayesian additivity. Can rationalise when the model is wrong.
- Need to be aware of the bias and variance trade-off in any density estimation technique.

The Solution: Hellinger Bayes

- Why use non-robust inference methodology for a decision problem?
- Appeal to general Bayesian updating to minimise a more robust divergence to the truth.
- Minimising a divergence equivalent to minimising a score [3].

Hellinger (H) divergence a robust approximation to KL in minimum discrepancy estimation [4].

$$d_H^2(g, f) = 1 - \int \sqrt{g(z)f(z)}dz \tag{4}$$

The H-divergence:

- Bounds Total-Variation both above and below.
- Bounded under contamination.
- Closeness in H-divergence means expected utility estimates will be absolutely close.

Bayesian updating using the score associated with the H-divergence (H-Bayes) is:

$$\pi(\theta|\mathbf{x}) \propto \exp\{w \sum_{i=1}^{n} (\frac{\sqrt{f_\theta(x_i)}}{\sqrt{g_n(x_i)}})\}\pi(\theta). \tag{5}$$

where $g_n(\cdot)$ is some non-parametric density estimate from the data.

- We contribute a foundational proposal for Bayesian updating using robust divergences that is valid in the M-open world.

Demonstration

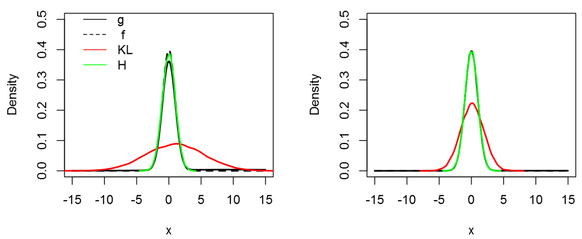

Figure 1: Posterior predictives: traditional Bayesian statistics (red) and H-Bayes (green), against the truth (black) and the approximating model (broken black). $\epsilon = 0.1$ (left) and $\epsilon = 0.01$ (right). A Kernel density estimate (KDE) estimates the true density.

E.g. 1: ϵ-contamination. Consider approximating genuine data generating function be g:

$$g = (1 - \epsilon)\mathcal{N}(0, 1) + \epsilon\mathcal{N}(\mu, \sigma^2), \tag{6}$$

with model f:

$$f = \mathcal{N}(0, 1). \tag{7}$$

E.g. 2: Over-dispersed Poisson. Approximating g:

$$g = \mathcal{NB}(s = \frac{1}{4}, \mu = 1), \tag{8}$$

with model f:

$$f = \text{Poi}(\lambda). \tag{9}$$

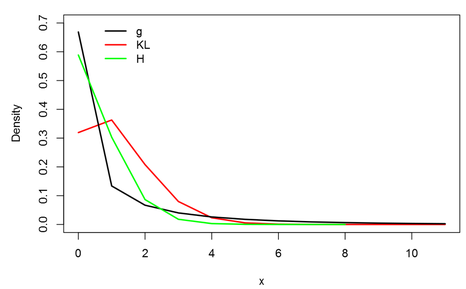

Figure 2: Posterior predictives produced by traditional Bayesian statistics (red) and H-Bayes (green) , against the truth (black). The empirical mass function estimates the true mass function.

References

[1] Bernardo, J, and Smith, A. (2001) "Bayesian Theory".
[2] Bissiri, P. Holmes, C. and Walker, S. (2016) JRSSB.
[3] Dawid, A.P. (2007) Annals of the Institute of Statistical Mathematics.
[4] Hooker, G and Vidyashankar, A. N. (2014) Test.

Next...

- Explore the power of H-Bayes updating under more general misspecification, targeting linear models.