Robust Bayesian Statistics

Jack Jewson: Supervised by Jim Smith (Warwick) and Chris Holmes (Oxford)

YRM: Tue 28 February 2017

Problem

"essentially, all models are wrong, but some are useful"
George E. P. Box

- In applied statistics a model, or group of models, are built based on expert judgements in order to provide the most realistic representation of a real world process.
- The goal of the statistician is then to produce the best inference or predictions.
- My goal is to look at how the statistician can make best use of these models
- We desire
  - Robust priors
  - Robust model selection
  - **Robust parameter updating**

Outline

- Look at the foundations of Bayesian statistics
- Demonstrate the lack of robustness exhibited by classical Bayesian statistics
- Propose an alternative to classical Bayesian statistics designed to improve the robustness.
- Demonstrate the advantages and limitations of this new method

Decision theory

- A decision maker (DM) is faced with making some decision $d \in \mathcal{D}$, that can be evaluated against some future unknown observation(s) $X \in \mathcal{X}$
- The consequences of taking decision d and observing outcome X are characterised by a loss function $\ell(d, X)$
- The DM wants to minimise their loss function (or equivalently maximise their utility)
- We will call this a decision problem

Savage Axioms

Savage axioms:

- A DM's beliefs can be represented by probabilities
- Their preferences by a utility function.
- Bayesian stats updates prior beliefs after data is observed, and finds the optimal decision as the one maximising the expected utility function (minimising the expected loss).

Bayesian Inference

Bernardo and Smith (2001):

- Bayesian inference is a decision problem for which the decision is to quote a probability belief distribution.
- The utility function associated with quoting a probability distribution must be proper and local
- The only proper local score function is

$$\ell(f(\cdot), \mathbf{x}) = -\sum_i \log(f(x_i))$$

However...

- The DMs true beliefs over future unknowns should encompass every possibly bit of information available at the time and may take a lifetime to write down.
- Instead beliefs over the data generating process are only ever going to be an approximation of the DM's true beliefs, usually using some convenient probability distribution and hopefully capturing the DM's important beliefs.
- If our model for the data will be misspecified compared to our true beliefs, is only considering inference a good idea?
- Should we be using inferential tools in a decision making context?

General Bayesian updating

Bissiri, Holmes, and Walker (2016):

- If we have a decision problem, a prior and some data, but no model, a Bayesian update must still be possible.
- The goal is to find the optimal decision (parametrised by θ) as:

$$\hat{\theta}(x) = \arg\min_{\theta(x)} \int \ell(\theta(x), x) dF_0(x)$$

- The posterior resulting from such an update should minimise a combination of the KL-divergence from the prior, and the expected loss of the data.
- The corresponding posterior is:

$$\pi(\theta|x) \propto \exp(-w \sum_{i=1}^{n} \ell(\theta, x_i))\pi(\theta)$$

- Where traditional Bayes builds a predictive model to 'best' approximate $F_0(x)$, general Bayes simply uses the empirical distribution of the observed data.

General Bayes: Pros and Cons

Pros:

- No longer reliant on a model and therefore robust to misspecifications
- Parameter no longer needs to index a probability distribution, can do Bayesian updating on anything, directly on a decision for example.

Cons:

- Don't have the structure of a model, i.e. the ability to use this posterior to produce a predictive
- No longer using Bayes rule for conditional probability so need to be careful to correctly set w such that our posterior maintains a probabilistic meaning.

General Bayesian updating: an example

Model free Bayesian clustering:

- Fixed number fo clusters K
- We directly do inference on the cluster allocations $C_i \in \{1, \dots, K\}$
- Minimise the squared error loss:

$$\ell(C, x) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_i^{(k)} - \bar{x}^{(k)})^2$$

where $\bar{x}^{(k)}$ is the mean (center) of cluster k.

Bayes as General Bayes

- If we consider doing general Bayes in order to do inference on a parameter θ then we want to use the local, proper, log-score:

$$\ell(\theta, x_i) = -\log(f(x_i; \theta))$$

- This results in the general Bayesian update equivalent to the traditional Bayesian update:

$$\pi(\theta|x) \propto \exp(-w \sum_{i=1}^{n} -\log(f(x_i; \theta)))\pi(\theta) = \prod_{i=1}^{n} f(x_i; \theta)\pi(\theta)$$

- An illustration that if a DM's utility function over a predictive distribution is characterised by the log-score then they should quote the Bayesian predictive distribution.

Divergence, Scores and Entropy

Dawid (2007):

- Considers inferential decision problems. Aiming to produce a predictive distribution minimising some loss function when used to predict future data.
- Defines:
  - a DM's expected score of quoting probability distribution f when g is true, $S(G,F) := \mathbb{E}_{X\sim G}[\ell(F,X)]$
  - the entropy of data generating distribution g, $H(G) = S(G,G) := \mathbb{E}_{X\sim G}[\ell(G,X)]$
  - the divergence between probability distributions, $D(G,F) := S(F,G) - H(G)$
- The divergence between two probability distributions, g and f, associated with loss function $\ell(\cdot,x)$ is the expected loss incurred for believing the data was distributed according to f when it was actually distributed according to g.

Minimising the KL-divergence

- If we consider attempting Bayesian inference then our divergence function of quoting f when g is true is:

$$D(G,F) = \int -g(x)\log(f(x))dx - \int -g(x)\log(g(x))dx = d_{KL}(G,F)$$

- So the traditional Bayesian predictive distribution minimises the KL-divergence between it and the truth.
- And the KL-divergence has the interpretation as the expected penalty for incorrectly predicting the true data generating probabilities
- In reality we don't know what the truth is so minimising the empirical log-score on the data set provides a proxy for minimising the KL-divergence.

Problems with KL

- The log-score focuses on the tails of the distribution by heavily penalising observations that would have been predicted with low probability $\lim_{x \to 0} \log(x) = \infty$.
- This is considered important for inference purposes but the tails might be less relevant for more general decision problems.
- This results in Bayesian statistics not being very robust to either small samples or misspecified models
- The Bayesian predictive focuses on capturing the tail behaviour of the observed data and this may result in poor predictive performance for the bulk of the data.

epsilon-contamination

- Consider producing a predictive distribution for modelling the height of people in China
- You collect a sample of 10,000 people
- In your sample was Chinese Basketball player Yao Ming

epsilon-contamination (cont.)

Consider the following toy example.

- Genuine data generating distribution.

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

- And lets say we have built a model that captures the majority of the population really well

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

- The KL divergence between g and f can be written as:

$$\begin{aligned} d_{KL}(g,f) = &\int_{-\infty}^{\infty}((1-\epsilon)h(x,0,1) + \epsilon h(x,\mu,\sigma^2(\mu))) \\ &\cdot \log(((1-\epsilon)h(x,0,1) + \epsilon h(x,\mu,\sigma^2(\mu))))dx \\ &- \int_{-\infty}^{\infty}((1-\epsilon)h(x,0,1) + \epsilon h(x,\mu,\sigma^2(\mu)))\log(h(x,0,1))dx \end{aligned}$$

where $h(x;\mu,\sigma)$ is the normal density, mean μ, variance σ^2

epsilon-contamination (cont.)

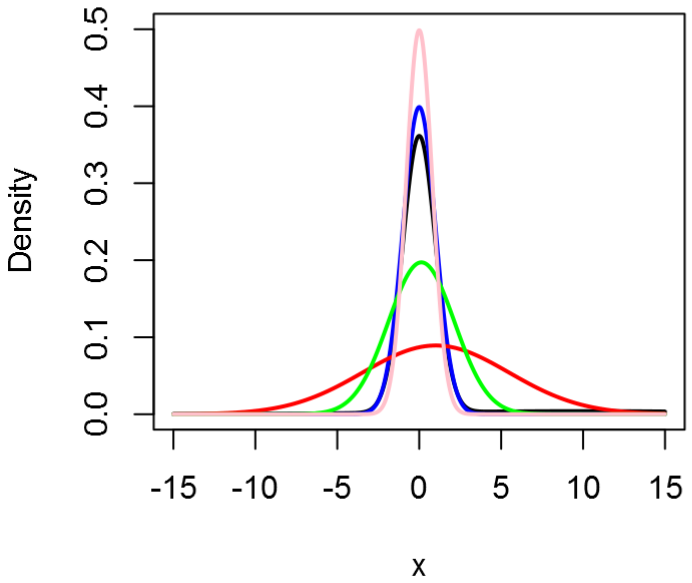
- The bottom equation reduces to:

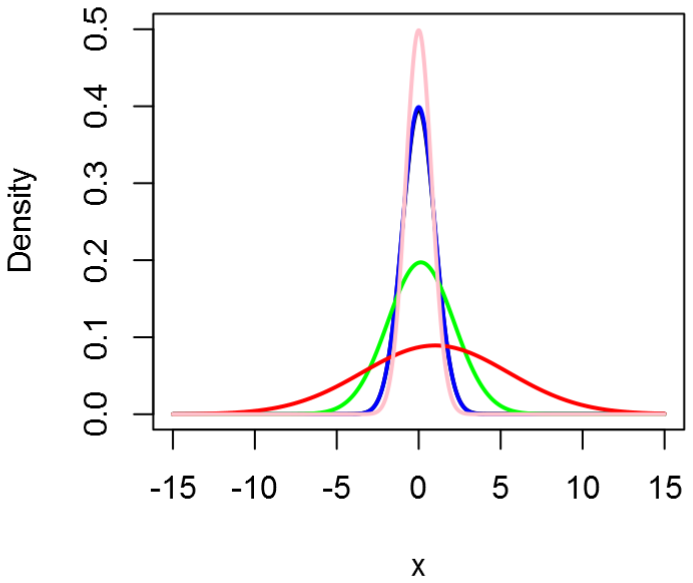
$$-[\log(\frac{1}{\sqrt{2\pi}}) - \frac{(1-\epsilon)}{2} - \frac{\epsilon(\sigma^2(\mu)+\mu^2)}{2}]$$

- Therefore as $\mu \to \infty$ and even if $\epsilon \to 0$, the KL-divergence will tend to ∞. Consider the following two illustration: $\mu = 10$, $\sigma = 10$ 1) $\epsilon = 0.1$, 2) $\epsilon = 0.01$

A little game

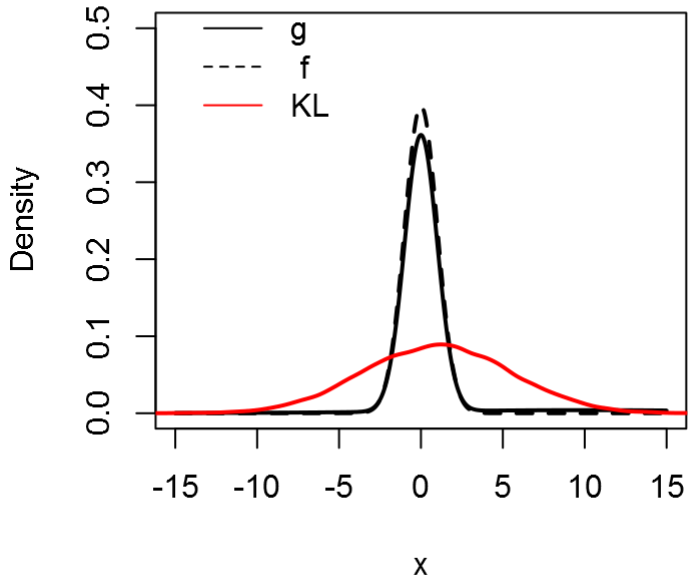
- Do you think you can guess which the KL minimiser (to the black line) is in each case?

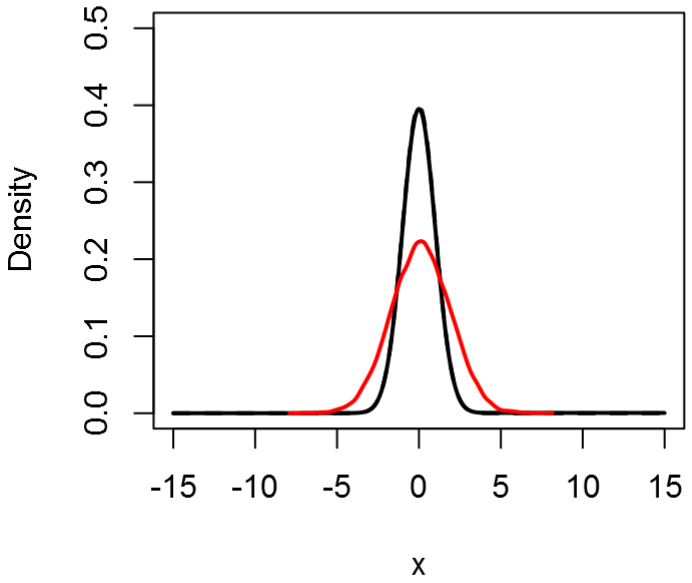




KL under misspecification

- We have a model that correctly captures the majority of the population but we have a misspecification in the tails.
- What does the KL minimiser look like? (all Bayesian inference done in *stan* see WRUG 02/03/17 for more)





Classical solutions

- Fit heavy tailed distributions e.g. Cauchy, Student's-t (see Berger et al. (1994))
- Frequentist M-estimators, for example a 'Huber-type' loss, can use general Bayes or pseudo-likelihoods (Greco, Racugno, and Ventura (2008)) for a Bayesian implementation

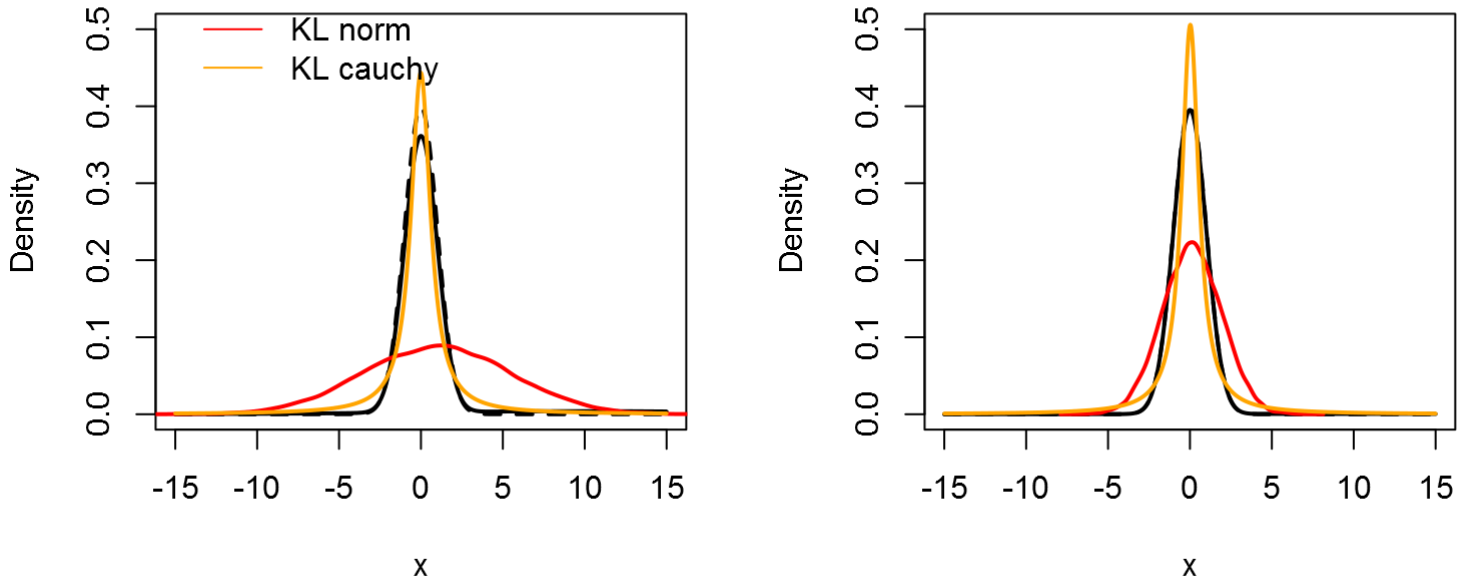
$$\ell(\theta, x) = (\theta - x)^2 \min(1, \frac{k}{(\theta - x)^2})$$

Or more recently

- Tempering the likelihood to slow the learning, 'safebayes' (Grünwald and Ommen (2014))

Heavy tails

- So lets try and fit a Cauchy distribution rather than a Gaussian distribution



Total-Variation

- Our idea: What if we use general Bayes to find the predictive that is closer to the truth in terms of some other divergence that is more robust than KL
- The gold standard would be close in Total-Variation:
  - Bounded under contamination
  - Minimising TV, minimise the chances/consequences of a Dutch-Book
  - Ensures accuracy of expected utility estimates

$$TV(F_0||F_\theta) = \int |f_0(x) - f_\theta(x)|dx$$

But… We are no longer have a strictly convex problem and the loss function is not differentiable.

Hellinger Divergence

- Consider the Hellinger divergence as a surrogate divergence for the Total-Variation
  - J. Smith (1995) show that the Hellinger divergence can bound the Total-Variation both above and below
  - Hooker and Vidyashankar (2014) implement in a Bayesian way justifying it as an approximation to the KL

$$H^2(F_0||F_\theta) = \int (\sqrt{f_0(x)} - \sqrt{f_\theta(x)})^2 dx = 1 - \int \sqrt{f_0(x) f_\theta(x)} dx$$

- We therefore propose the following Bayesian update:

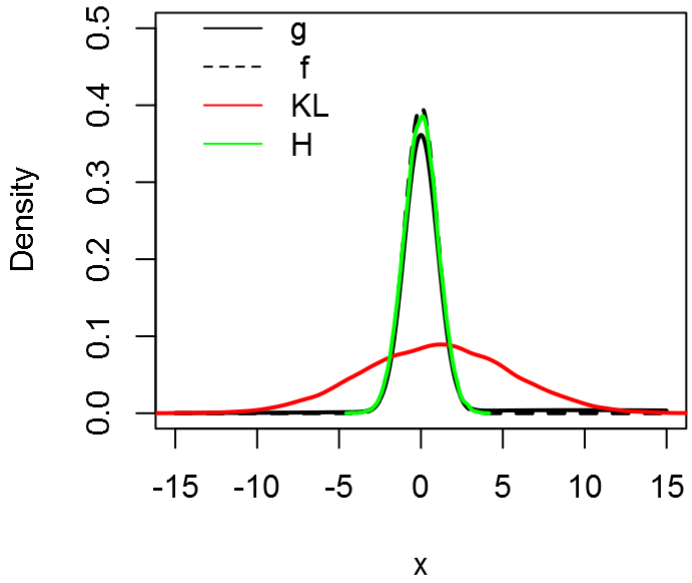
$$\pi(\theta|X) \propto \exp(w(\frac{\sqrt{f_\theta(x)}}{\sqrt{g_n(x)}}))\pi(\theta)$$

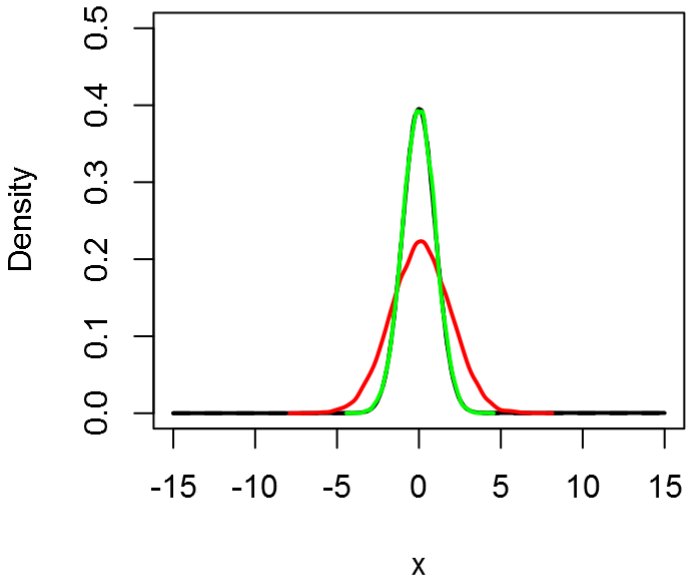
where $g_n(x)$ is some density estimate aimed to approximate the true sampling density $f_0(x)$

Bayesian coherence and the likelihood principle

- The likelihood principle says: The likelihood is sufficient for the data
- Bayesian coherence says: Our posterior looks the same if we update in one go or sequentially
- Locality: demands our models are only scored based on the predicted probabilities of observed data
- All three are abandoned under misspecification, due to the introduction of an empirical estimate:
  - If the model is wrong the likelihood is no longer sufficient for the data, there is more information in the data
  - Follows from abandoning the likelihood principle. There is more information in the data when viewed together
  - Don't want to score a model solely on predictions we know are not correct

The Hellinger under contamination





What must a robust update satisfy?

Demands:

- Proper: It must converge to the correct minimiser and this be the truth if the model is correct.
- Posteriors must have probabilistic meaning: Bayesian stats correctly updates prior uncertainty to posterior uncertainty in light of data using Bayes rule, we no longer have Bayes rule so need to make sure our posterior uncertainty means something. i.e. that the posterior is extracting the 'correct' amount of information in the data
- Robust: Need to be able to show more general robustness properties.

Issues:

- To ensure we get convergence to the desired minimiser the the density estimate must be carefully chosen. For small samples KDEs have biases! How will we come up with a density estimate in a predictor response scenario like linear regression?
- Need to think carefully about how to set the learning rate. Above $w = 1$ arbitrarily. Shouldn't be able to learn faster than a Bayesian armed with the truth.

Further work

- Hellinger not the only divergence, there are many others to be explored.
- Need to explore performance in more involved examples - linear regression for treatment assignment
- Updating is not the only element of the Bayesian machine.
  - Robust prior specification
  - Robust model selection

Being robust when you update is useless if your starting model is not.

References

Berger, James O, Elías Moreno, Luis Raul Pericchi, M Jesús Bayarri, José M Bernardo, Juan A Cano, Julián De la Horra, et al. 1994. “An Overview of Robust Bayesian Analysis.” *Test* 3 (1). Springer: 5–124.

Bernardo, José M, and Adrian FM Smith. 2001. “Bayesian Theory.” IOP Publishing.

Bissiri, PG, CC Holmes, and Stephen G Walker. 2016. “A General Framework for Updating Belief Distributions.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Wiley Online Library.

Dawid, A Philip. 2007. “The Geometry of Proper Scoring Rules.” *Annals of the Institute of Statistical Mathematics* 59 (1). Springer: 77–93.

Greco, Luca, Walter Racugno, and Laura Ventura. 2008. “Robust Likelihood Functions in Bayesian Inference.” *Journal of Statistical Planning and Inference* 138 (5). Elsevier: 1258–70.

Grünwald, Peter, and Thijs van Ommen. 2014. “Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It.” *ArXiv Preprint ArXiv:1412.3730*.

Hooker, Giles, and Anand N Vidyashankar. 2014. “Bayesian Model Robustness via Disparities.” *Test* 23 (3). Springer: 556–84.

Smith, JQ. 1995. “Bayesian Approximations and the Hellinger Metric.”