Principled Robust Bayesian Updating

with Applications to Online Changepoint Detection

Jack Jewson, University of Warwick

BAYSM, Warwick, 02/07/18

supervised by Jim Q. Smith and Chris Holmes (Oxford), in collaboration with Jeremias Knoblauch and Theo Damoulas

The M-open world

- M-closed world: There exists a parameter θ_0 such that the data $X \sim f(\cdot; \theta_0)$
- M-open world:

> "All models are wrong but some are useful"
>
> G. E. P. Box

- The model is misspecified vs the sample distribution of the data.
- Cannot learn θ_0 generating the data.
- Define parameter of interest by defining **divergence** between model and sample distribution of the data (Walker, 2013) (JSPI).

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 sample distribution of x.
- The traditional Bayesian builds a belief model to approximate $G(x)$.
- But this belief model will inevitably be misspecified - M-open world.

General Bayesian Updating

- Given a prior and a loss function, an updating of beliefs in light of data must be possible even without a model.
- In such a scenario the General Bayesian's posterior beliefs (Bissiri, Holmes and Walker, 2016) (JRSSB) 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:

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

Bayes as General Bayes

If $\ell(\theta, x) = -\log(f(x;\theta))$ then the general Bayesian update recovers Bayes rule:

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

- Bayesian updating is learning about the parameter which minimises the **KL-divergence** to the sample distribution.
- But as $f(x;\theta) \to 0$, $-\log(f(x;\theta)) \to \infty$.
- Results in an (implicit) desire to correctly capture the **tail behaviour** of the underlying process.
- In order conduct **principled inference** in the M-open world, the DM is currently forced to worry about how **robust** the tails of their model are.

ϵ-contamination

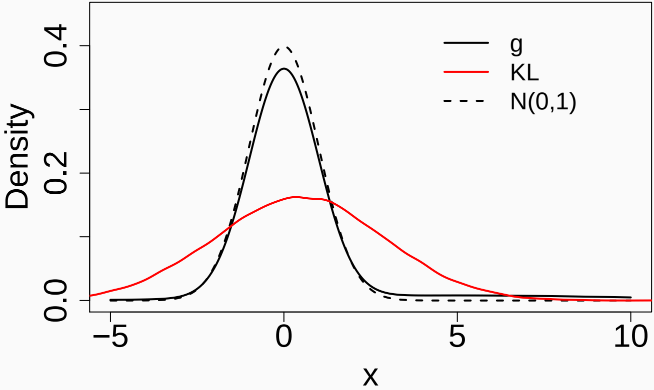

Figure 1 – Posterior predictive distribution fitting $\mathcal{N}(\mu, \sigma^2)$ to $g = 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(5, 5^2)$ using the traditional Bayesian updating (KL-Bayes).

A Principled Alternative

- Each divergence $d(\cdot, \cdot)$ has a corresponding loss function $\ell_d(\cdot, \cdot)$
- General Bayesian updating therefore allows for principled belief updating for parameters minimising divergences **other than KL-divergence** (Jewson, Smith and Holmes, 2018) (Entropy)

$$\pi^{(d)}(\theta|\mathrm{x}) \propto \pi^{(d)}(\theta) \exp\left(-\sum_{i=1}^{n} \ell_d(x_i, f(\cdot;\theta))\right). \tag{4}$$

- Not a pseudo or approximate posterior as previously thought (Hooker and Vidyashankar, 2014 (Test), Ghosh and Basu, 2016 (AISM)) .
- $w = 1$ as doing model based inference with a well-defined divergence.

The divergence as a subjective judgement

- Principled justification allows the divergence to become a **subjective judgement** alongside prior and model.
- Represents how strongly you believe in your model (especially its tails).
- Decouples belief elicitation and robustness.
- Decision theoretic reasons for Total Variation (TV), Hellinger (Hell) or α-divergences, but these require a density estimate.
- Alternatively the β-divergence with loss

$$\ell_\beta(\theta, x) = \frac{1}{1+\beta}\int_{\mathcal{Y}} f(y;\theta)^{\beta+1}dy - \frac{1}{\beta}f(x;\theta)^\beta. \tag{5}$$

ϵ-contamination

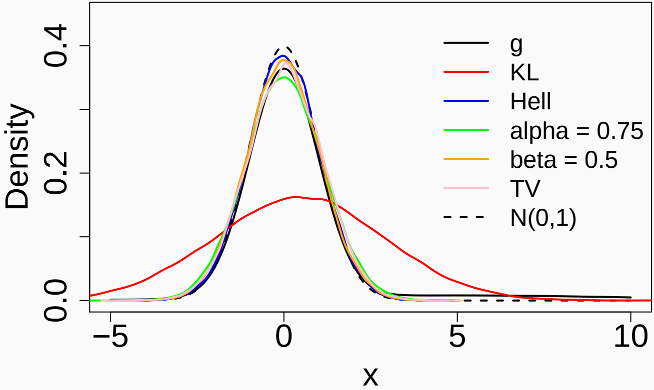

Figure 2 – Posterior predictive distributions fitting $\mathcal{N}(\mu, \sigma^2)$ to $g = 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(5,5^2)$ using the KL-Bayes , Hell-Bayes, TV-Bayes , alpha-Bayes ($\alpha = 0.75$) and beta-Bayes ($\alpha = 0.5$).

Influence under the β-divergence

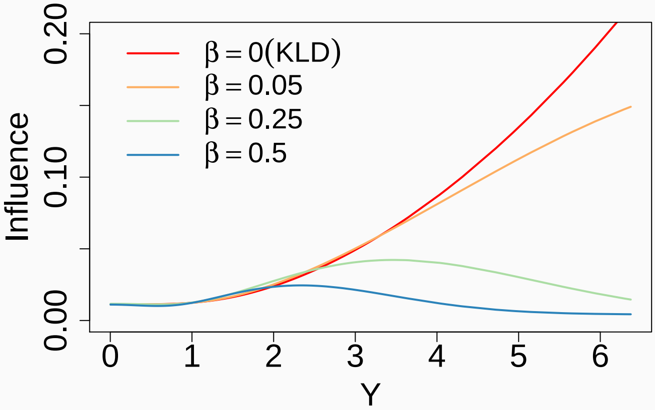

Figure 3 – The influence (Kurtek and Bharath, 2015) (Biometrika) of removing one of 1000 observations from a $t(4)$ distribution when fitting a $\mathcal{N}(\mu, \sigma^2)$ under the beta-Bayes for different values of β.

Bayesian On-line Changepoint Detection (BOCPD)

e.g. Knoblauch and Damoulas (2018) (ICML)

- Quantify change-point uncertainty with a **run length posterior**

$$\begin{aligned} \pi^{(KL)}(r_t = t - l | x_{1:t}) \propto & p(r_l = 0, x_{1:l}) \prod_{i=l}^{t} p(r_i | r_{i-1}) \prod_{i=l}^{t} p^{(KL)}(x_i | x_{l:i-1}) \\ = & \pi_0(r_t = t - l) \exp\left(- \sum_{i=l}^{t} - \log\left(p^{(KL)}(x_i | x_{l:i-1}) \right) \right) \end{aligned} \tag{6}$$

- Uses the **predictive density** of next observation as the run length likelihood.
- **Outliers** have low predictive density and cause **spurious CPs**

Robust BOCPD

Knoblauch, Jewson and Damoulas (2018) (arxiv)

- Maintains full and principled uncertainty quantification with robust run length posterior

$$\pi^{(\beta)}(r_t = t - l | x_{1:t}) \propto \pi_0(r_t = t - l) \exp\left(-\sum_{i=l}^{t} \ell_\beta(r_i, x_i)\right).$$

- Can set hyperparameters such that **one observation** alone cannot declare a CP.
- Propose a **structured (quasi-conjugate) variational inference** routine to conduct high-dimensional parameter posterior inference using the β-divergence on-line.
- **Initialise** β to give maximum influence to regions where data is *a priori* expected to arrive and **update** β on-line using a higher level loss.

Synthetic example

Five-dimensional Vector Autoregression (VAR) with one dimension injected with t_4-noise.

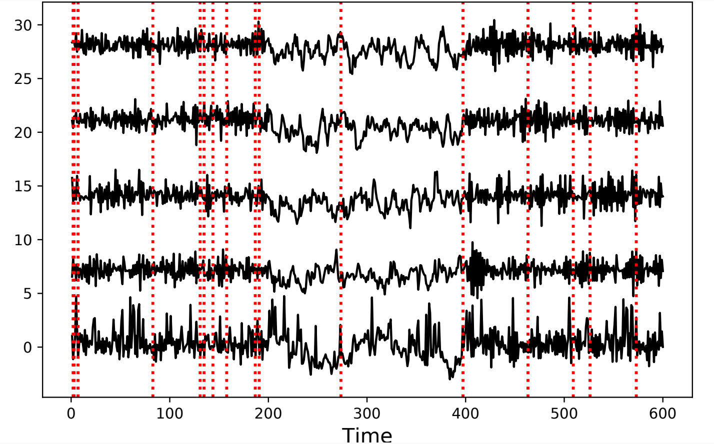

Figure 4 – Maximum A Posteriori (MAP) CPs of standard BOCPD shown as dashed vertical lines. True CPs at $t = 200, 400$.

Synthetic example

Five-dimensional Vector Autoregression (VAR) with one dimension injected with t_4-noise.

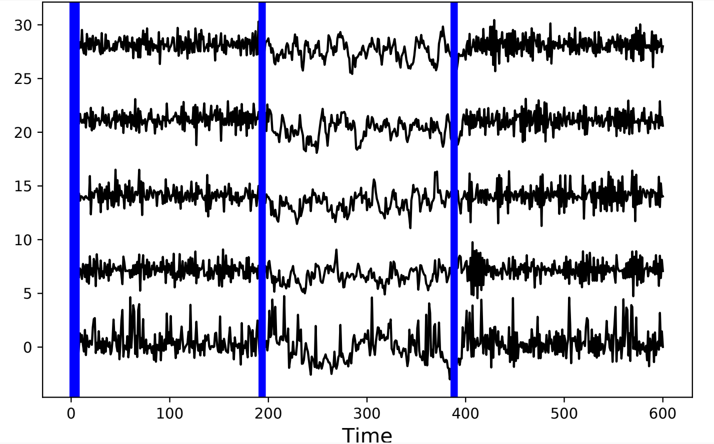

Figure 5 – Maximum A Posteriori (MAP) CPs of robust (standard) BOCPD shown as solid (dashed) vertical lines. True CPs at $t = 200, 400$.

London Air Pollution

Dataset recording Nitrogen Oxide levels across 29 stations in London modelled using three spatially structured Bayesian VARs.

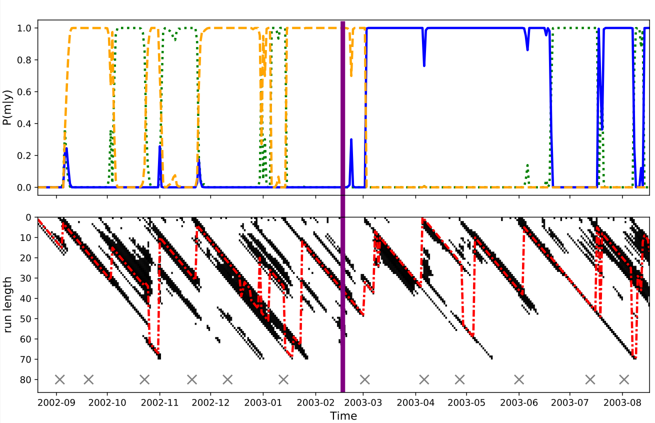

Figure 6 – most likely run-lengths for standard BOCPD. Also marked are the congestion charge introduction, 17/02/2003 (solid vertical line) and the MAP segmentations (crosses).

Dataset recording Nitrogen Oxide levels across 29 stations in London modelled using three spatially structured Bayesian VARs.

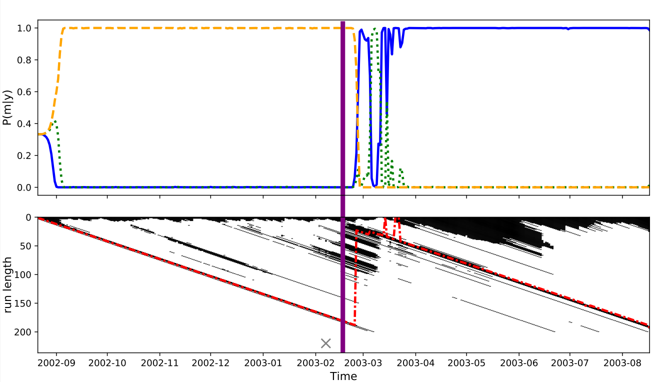

Figure 7 – most likely run-lengths for robust BOCPD. Also marked are the congestion charge introduction, 17/02/2003 (solid vertical line) and the MAP segmentations (crosses).

Further Work

- Theory/Axioms why you should not update beliefs using the KL-divergence.
- Formalise quasi-conjugate variational inference for further families, estimating equations, guarantees on performance.
- Robust Bayesian model selection, run-length posterior similar to model selection posterior.
- Other application areas - open to ideas!