Robust Bayesian Updating

Jack Jewson, Jim Q. Smith and Chris Holmes (Oxford) collaborating with Jeremias Knoblauch and Theo Damoulas

M-open Inference

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

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

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

where $G(x)$ is the sample distribution of x.
- The traditional Bayesian builds a belief model to approximate $G(x)$.

Without a model, 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:

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

Simple Demonstration

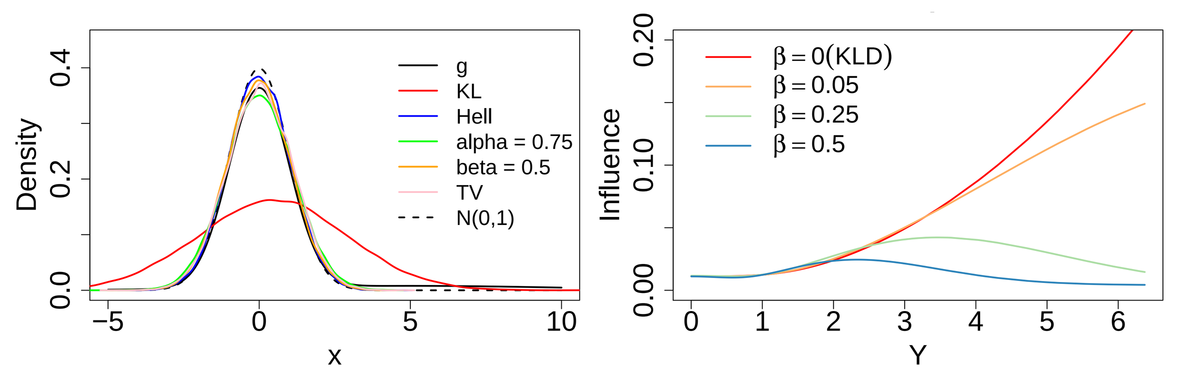

Figure 4: Left: Posterior predictive distributions fitting $\mathcal{N}(\mu, \sigma^2)$ to $g = 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(5, 5^2)$ using the KL-Bayes , Hell-Bayes, TV-Bayes , alpha-Bayes ($\alpha = 0.75$) and beta-Bayes ($\alpha = 0.5$). Right: The influence (Kurtek and Bharath, 2015) (Biometrika) 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.

Bayes as General Bayes

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

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

- Bayesian updating is learning about the parameter which minimises the **KL-divergence** to the sample distribution of the data.
- 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 to conduct **principled inference**.

A Principled Alternative

- Each divergence $d(\cdot,\cdot)$ has a corresponding loss function $\ell_d(\cdot,\cdot)$
- Equation (2) allows for principled belief updating for parameter minimising divergences **other than KL-divergence** (Jewson, Smith and Holmes, 2018) (Entropy)

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

- Not a pseudo or approximate posterior as previously thought.
- $w = 1$ as doing model based inference with a well-defined divergence.
- 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 alpha-divergences, but these require a density estimate.
- Alternatively the β-divergence with loss

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

Robust Bayesian Online Changepoint Detection (BOCPD)

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

- Detects **Changepoints** (CPs) online providing full uncertainty quantification.
- Combines a **run-length** (time since last CP) posterior and parameter posterior within segment.
- Use **predictive density** of next observation as the run length likelihood.
- **Outliers** have low predictive density and cause **spurious CPs**.
- Efficient recursion to update posterior online.

Robust BOCPD (Knoblauch, Jewson and Damoulas (2018) (arxiv))

- Maintains full and principled uncertainty quantification.
- **Robustify run-length** posterior using the β-**divergence score** in place of the log-score.
- Can set hyperparameters such that **one observation** alone cannot declare a CP.
- Also use the β-divergence for the parameter posterior.
- Propose a **structured (quasi-conjugate) variational inference** routine to conduct high-dimensional inference for the β-Bayes online.
- **Initialise** β to give maximum influence to regions where data is *a priori* expected to arrive.
- **Update** β online using a higher level loss.
- Could possibly be extended to **Robust Bayesian model selection** using a loss function on the prior predictive.

Synthetic example

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

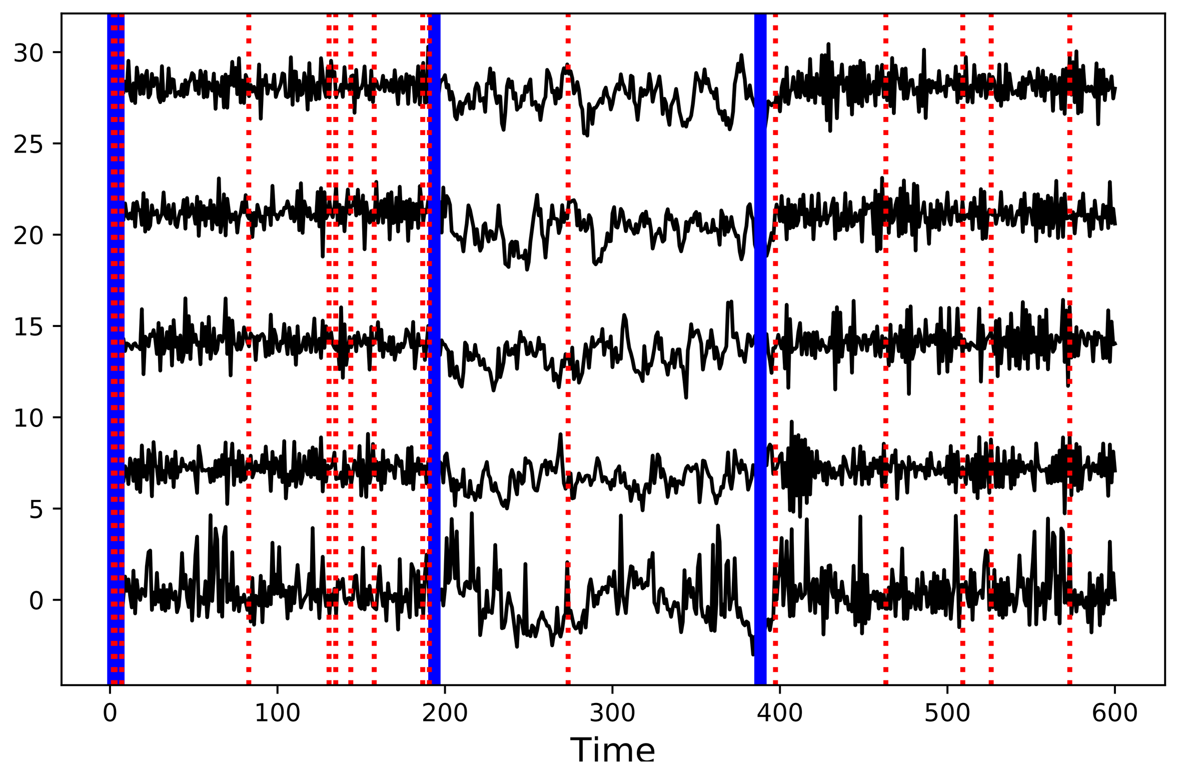

Figure 1: Maximum A Posteriori (MAP) CPs of robust (standard) BOCPD shown as solid (dashed) vertical lines. True CPs at $t = 200, 400$. In **high dimensions** it becomes increasingly likely that the model's tails are misspecified in at least one dimension.

'well-log' dataset

Univariate data seeking to detect changes in rock strata.

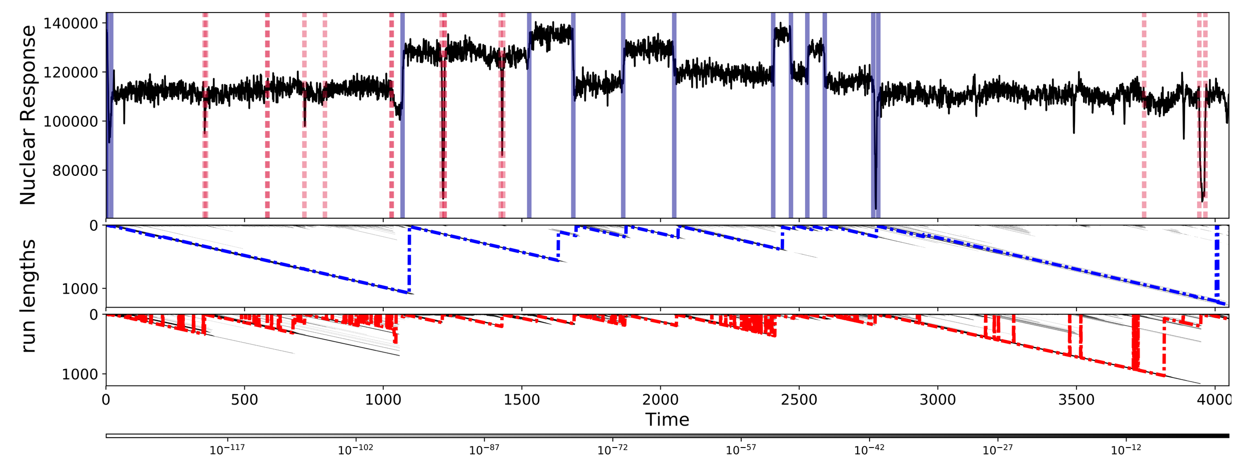

Figure 2: Maximum A Posteriori (MAP) segmentation and run-length distributions of the well-log data. Robust segmentation depicted using solid lines, CPs additionally declared under standard BOCPD with dashed lines. The corresponding run-length distributions for robust (middle) and standard (bottom) BOCPD are shown in greyscale. The most likely run-lengths are dashed.

London Air Pollution

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

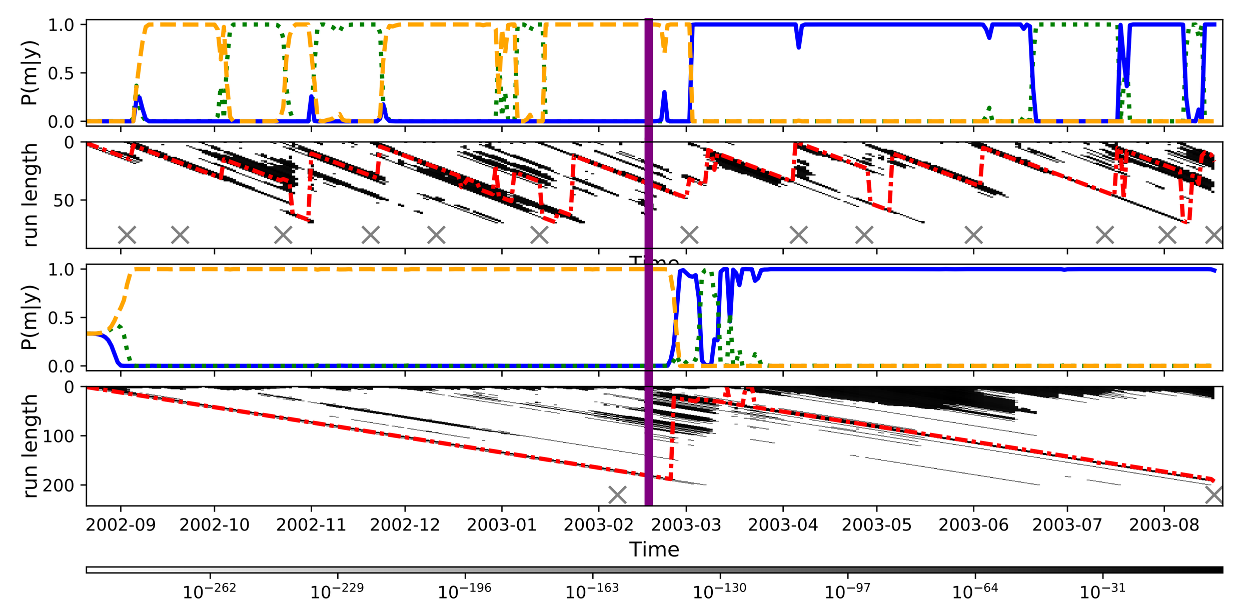

Figure 3: On-line model posteriors for three different VAR models (solid, dashed, dotted) and run-length distributions in greyscale with most likely run-lengths for standard (top two panels) and robust (bottom two panels) BOCPD. Also marked are the congestion charge introduction, 17/02/2003 (solid vertical line) and the MAP segmentations (crosses).