Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences

Jeremias Knoblauch[1,3] j.knoblauch@warwick.ac.uk
Jack Jewson[1], j.e.jewson@warwick.ac.uk
Theodoros Damoulas[1,2,3] t.damoulas@warwick.ac.uk

[1]University of Warwick, Department of Statistics
[2]University of Warwick, Department of Computer Science
[3]The Alan Turing Institute

The Alan Turing Institute | Lloyd's Register Foundation | OxWaSP | WARWICK THE UNIVERSITY OF WARWICK

The Problem

Inference in non-stationary data through Bayesian On-line Changepoint Detection (BOCPD) fails for high dimensions and outliers.

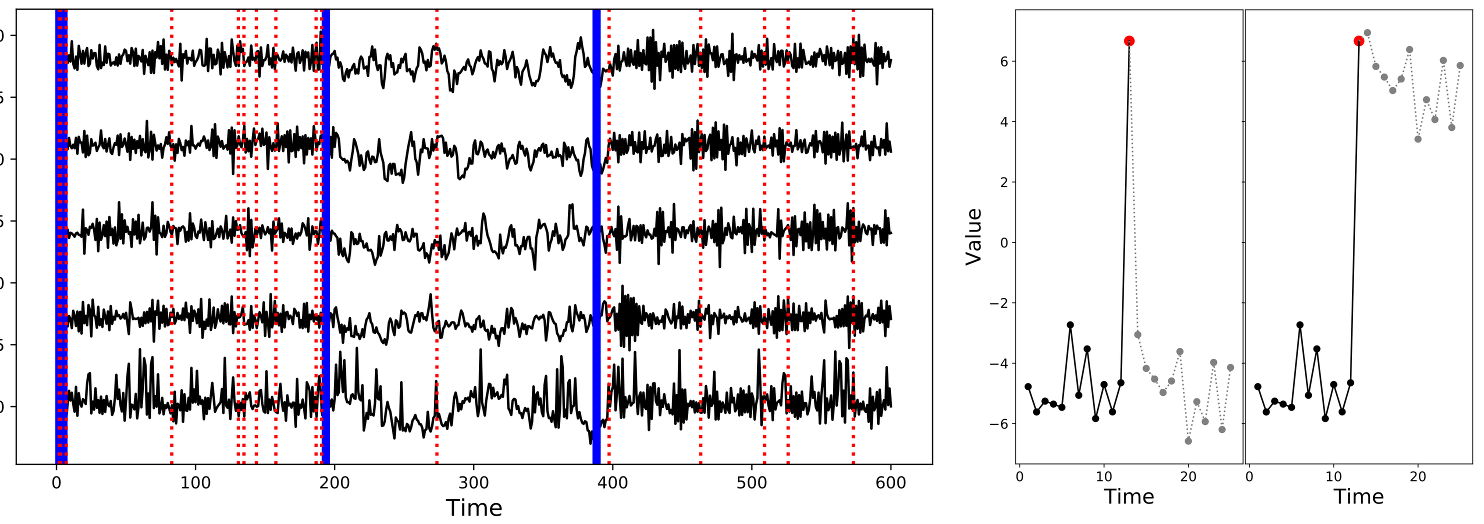

Figure 1: **Left: Standard** BOCPD on 5-dimensional AR(1) with 3 **true** changepoints. **Right:** BOCPD's sequential inference cannot distinguish outliers and changepoints.

The Solution

Jewson, Smith and Holmes (2018) introduce generalized Bayes Theorems for optimal belief updating under different divergences. For model m with density f_m, this takes the form

$$\pi_m^D(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \propto \pi_m(\theta)\exp\left\{-\textstyle\sum_{i=t-r_t}^{t}\ell^D(\boldsymbol{\theta}_m|\boldsymbol{y}_i)\right\} \quad (1)$$

$$\ell^{\text{KLD}}(\boldsymbol{\theta}_m|\boldsymbol{y}_t) = -\log\left(f_m(\boldsymbol{y}_t|\boldsymbol{\theta}_m)\right) \quad (2)$$

$$\ell^{\beta}(\boldsymbol{\theta}_m|\boldsymbol{y}_t) = -\left(\frac{1}{\beta_p}f_m(\boldsymbol{y}_t|\boldsymbol{\theta}_m)^{\beta_p} - \frac{1}{1+\beta_p}\int_{\mathcal{Y}} f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)^{1+\beta_p}d\boldsymbol{z}\right) \quad (3)$$

D = **Kullback-Leibler Divergence (KLD)** recovers the traditional Bayes Theorem; setting $D = \beta$ yields robust updates via the **β-Divergence (βD)**.

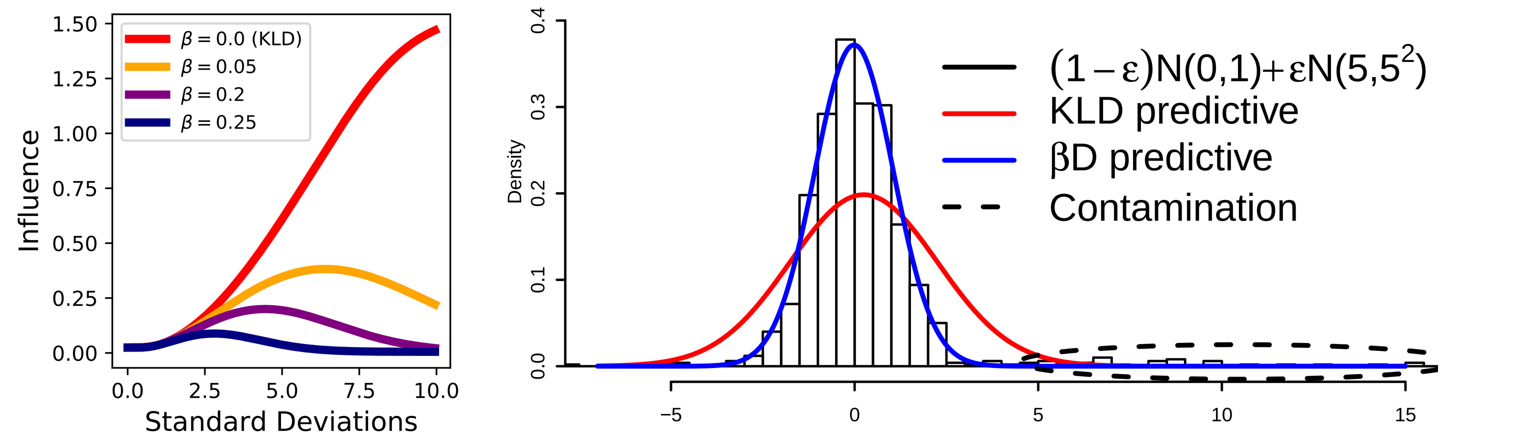

Figure 2: **Left:** Influence functions for different β and the KLD. **Right:** $\epsilon = 0.05$ contaminated data and its **KLD** and **βD** ($\beta = 0.5$) posterior predictive distributions

The Result

Following Knoblauch & Damoulas (2018), BOCPD is written as

$$r_t|r_{t-1} \sim H(r_t, r_{t-1}) \qquad m_t|\{r_t = 0\} \sim q(m_t) \quad (4)$$

$$\boldsymbol{\theta}_{m_t} \sim \pi_{m_t}(\boldsymbol{\theta}_{m_t}) \qquad \boldsymbol{y}_t \sim f_{m_t}(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t}) \quad (5)$$

which enables efficient recursive and doubly robust inference via

$$f_{m_t}^{\beta_p}(\boldsymbol{y}_t|\boldsymbol{y}_{(t-r_t):(t-1)}, r_t) = \int_{\Theta} f_{m_t}(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t})\pi_m^{\beta_p}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t})d\boldsymbol{\theta}_{m_t} \quad (6)$$

$$p^{\beta_{\text{rlm}}}(\boldsymbol{y}_{1:t}, r_t, m_t) \propto \sum_{m_{t-1}, r_{t-1}}\Big\{e^{-\ell^{\beta_{\text{rlm}}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_{t-1}):(t-1)})}p^{\beta_{\text{rlm}}}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \; H(r_t, r_{t-1})q^{\beta_{\text{rlm}}}(m_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1})\Big\}. \quad (7)$$

Reducing FDR to 0% on real world data

Outliers in the well log data are usually excluded to avoid outliers being mislabelled as changepoints, but robust BOCPD achieves 0% FDR without such preprocessing.

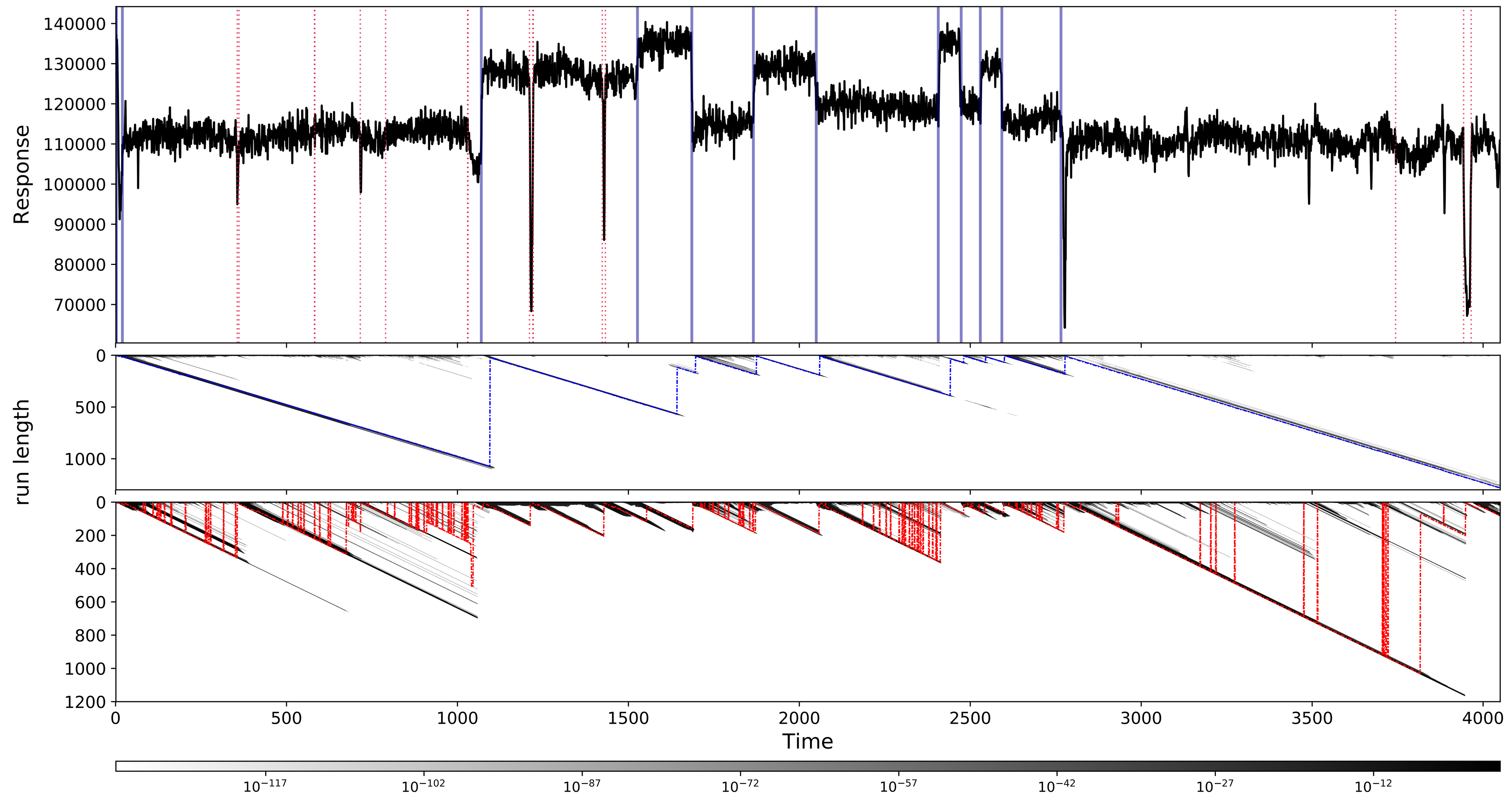

Figure 3: **Top:** well log data and changepoints found with **robust** BOCPD as solid lines. Additional changepoints found with **standard** BOCPD as dotted lines. **Middle: Robust** run-length posterior in grayscale, with emphasized maximum. **Bottom: Standard** run-length posterior in grayscale, with emphasized maximum.

London's Congestion Charge

BOCPD on 29 Air Pollution sensors in London. The robust version finds the Congestion Charge introduction date while the moderate problem dimension renders the standard version fragile.

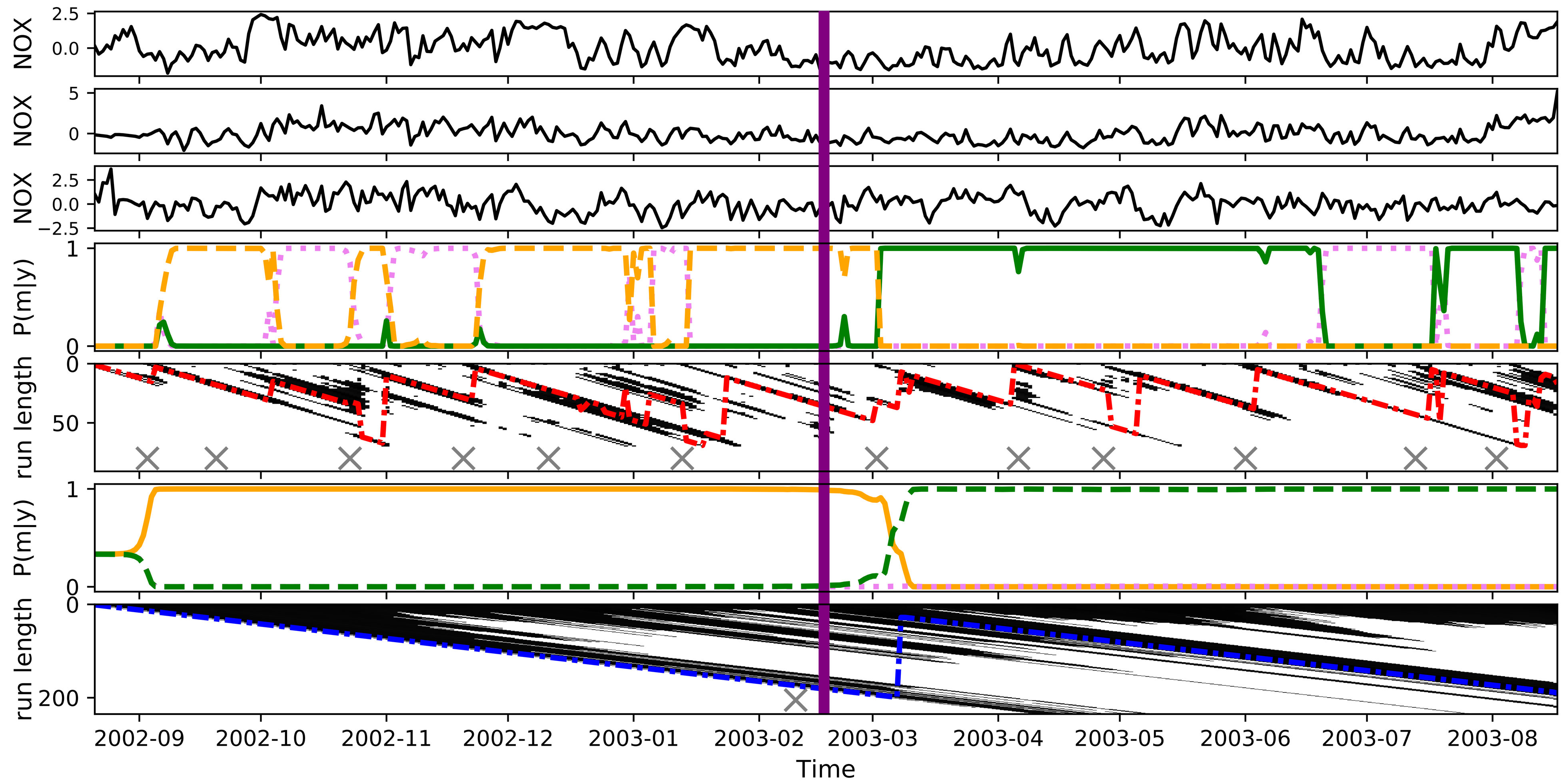

Figure 4: **All Panels:** Introduction of **London's Congestion Charge** as vertical line. **Panels 1–3:** Nitrogen Oxide measurements across London for 3/29 analyzed stations. **Panels 4–5:** On-line model posterior and run-length posteriors of **standard** BOCPD with detected changepoints marked as crosses ($\times$) and emphasized maximum run-length. **Panels 6–7:** On-line model posterior and run-length posteriors of **robust** BOCPD with detected changepoints marked as crosses ($\times$) and emphasized maximum run-length.

Thm. 1: Robustness Guarantee

Q: Why not simply use Student's t errors instead?
(A) Can not robustify asymmetric/discrete/... problems;
(B) Models outliers as *part* of the DGP;
(C) Provides no robustness in changepoint posteriors (!)
The β-D trivially solves (A–B). Thm. 1 shows (C) is also not an issue as

$$\frac{p(r_{t+1} = r+1|\boldsymbol{y}_{1:t+1}, r_t = r, m_t)}{p(r_{t+1} = 0|\boldsymbol{y}_{1:t+1}, r_t = r, m_t)} \geq 1. \quad (8)$$

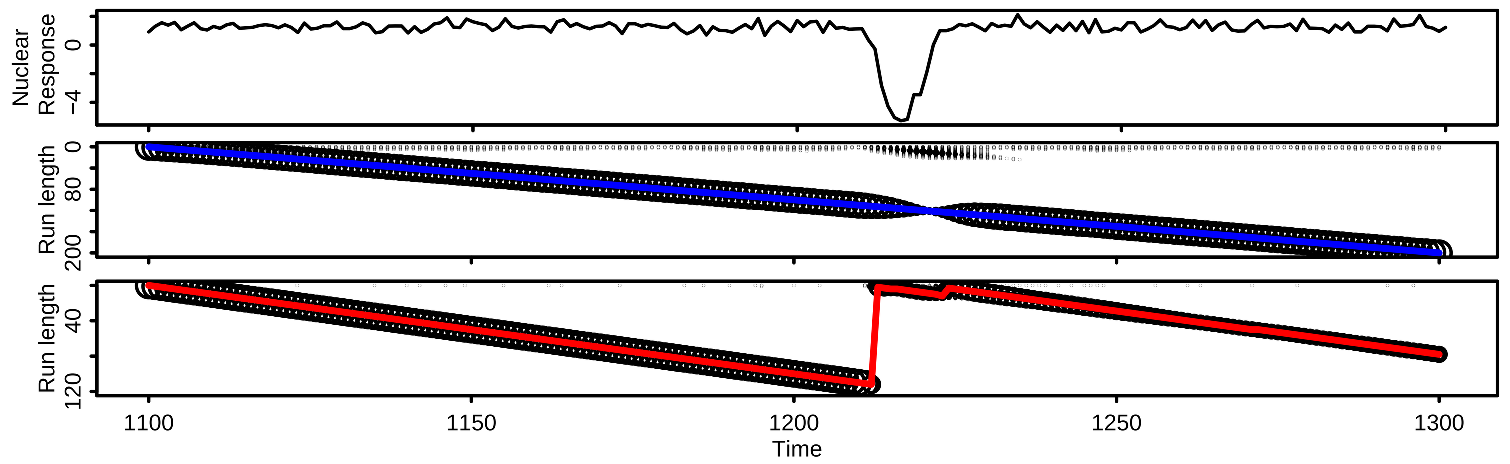

Figure 5: **Top**: 200 observations from the well log data. **Middle**: Gaussian **βD** run-length posterior. **Bottom**: Student's t_5 **KLD** run-length posterior.

Thm. 2: Efficient Approximation

One gets a closed form ELBO for the structural variational approximation

$$\widehat{\pi}_m^{\beta_p}(\boldsymbol{\theta}_m) = \underset{\pi_m^{\text{KLD}}(\boldsymbol{\theta}_m)}{\operatorname{argmin}}\left\{\text{KL}\left(\pi_m^{\text{KLD}}(\boldsymbol{\theta}_m)\,\middle\|\,\pi_m^{\beta_p}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t})\right)\right\}. \quad (9)$$

This means it is solvable with standard optimizers. $\widehat{\pi}_m^{\beta_p}(\boldsymbol{\theta}_m)$ is also attractive as it (I) is exact as $\beta_p \to 0$ and (II) captures parameter dependence.

Optimal choice of β

β is initialized to maximize influence of observations at a prespecified point and optimized on-line to minimize prediction error:

$$\beta_t = \beta_{t-1} + \gamma_t \cdot \nabla_{\beta_{t-1}} L(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t(\beta_{t-1})) \quad (10)$$

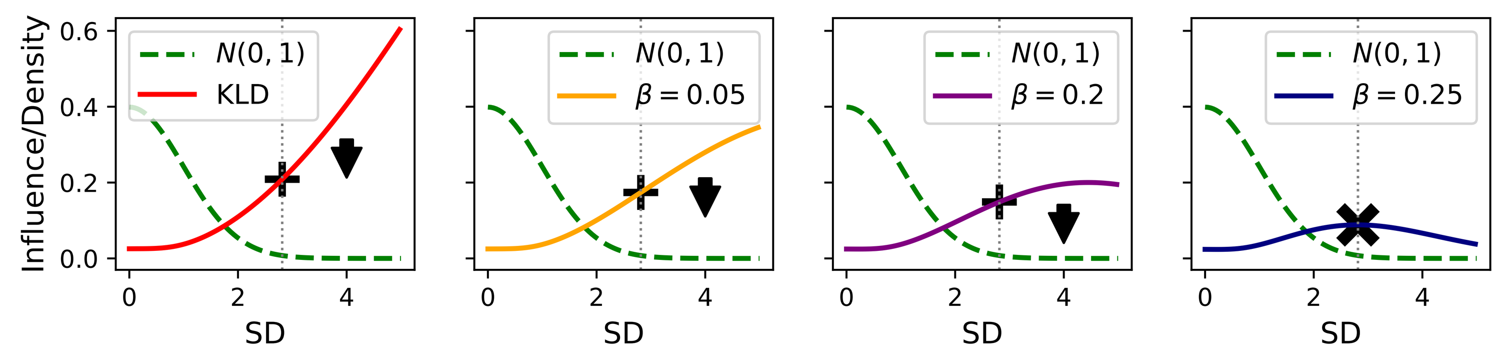

Figure 6: Initializing β by choosing a point of **maximum influence**

Key References

Adams, R.P. & JC. MacKay, D. J. C. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Bissiri, P.G., Holmes, C. C. & Walker, S. G. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5), pp.1103-1130, 2016.

Fearnhead, P. & Liu, Z. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B*, 69 (4), pp. 589–605, 2007.

Jewson, J., Smith, J.Q. & Holmes, C. Principles of Bayesian Inference Using General Divergence Criteria. *Entropy*, 20(6), pp.442–467, 2018.

Knoblauch, J. & Damoulas, T. Spatiotemporal Bayesian On-line Changepoint Detection with Model Selection. *Proceedings of the 35th International Conference on Machine Learning*, 2018.

CODE | PAPER | VIDEO | Montréal 2018