

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

The Alan Turing Institute

Lloyd's Register Foundation

OxWaSP

WARWICK
THE UNIVERSITY OF WARWICK

Jeremias Knoblauch^{1,3}
Jack Jewson¹,
Theodoros Damoulas^{1,2,3}

j.knoblauch@warwick.ac.uk
j.e.jewson@warwick.ac.uk
t.damoulas@warwick.ac.uk

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

THE PROBLEM

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

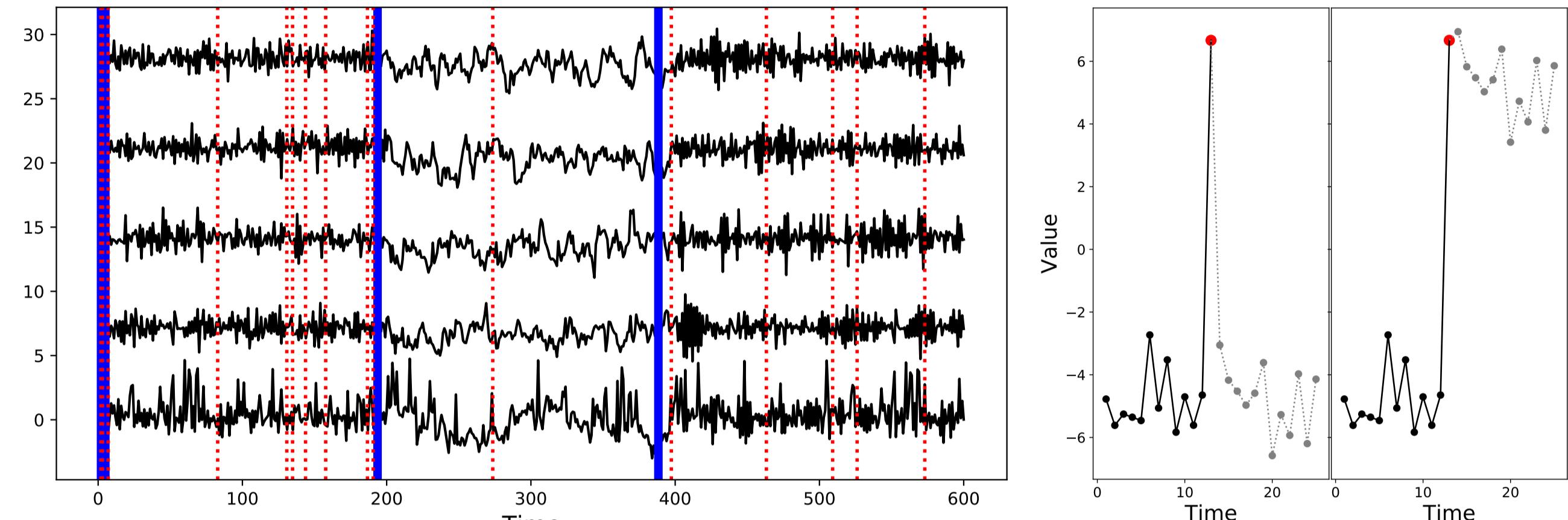


Figure 1: Left: Standard BOCPD on 5-dimensional AR(1) with 3 true changepoints. Right: BOCPD's sequential inference cannot distinguish outliers and changepoints.

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

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

$$\ell^{\text{KLD}}(\theta_m | \mathbf{y}_t) = -\log(f_m(\mathbf{y}_t | \theta_m)) \quad (2)$$

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

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

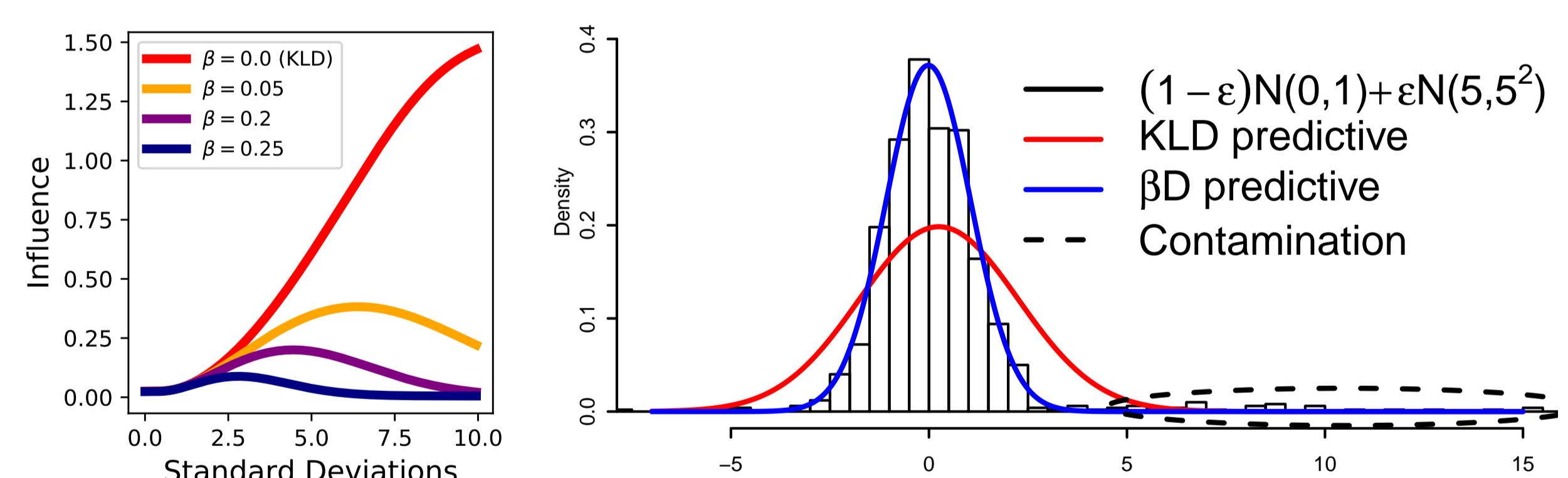


Figure 2: Left: Influence functions for different β and the KLD. Right: $\epsilon = 0.05$ contaminated data and its KLD and β D ($\beta = 0.5$) posterior predictive distributions

THE RESULT

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

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

$$\theta_{m_t} \sim \pi_{m_t}(\theta_{m_t}) \quad \mathbf{y}_t \sim f_{m_t}(\mathbf{y}_t | \theta_{m_t}) \quad (5)$$

which enables efficient recursive and doubly robust inference via

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

$$p^{\beta_{\text{rim}}}(\mathbf{y}_{1:t}, r_t, m_t) \propto \sum_{m_{t-1}, r_{t-1}} \frac{\{e^{-\ell^{\beta_{\text{rim}}}(\theta_m | \mathbf{y}_{(t-r_{t-1}):t-1})} p^{\beta_{\text{rim}}}(\mathbf{y}_{1:(t-1)}, r_{t-1}, m_{t-1})\}}{H(r_t, r_{t-1}) q^{\beta_{\text{rim}}}(m_t | \mathbf{y}_{(t-1)}, r_{t-1}, m_{t-1})} \quad (7)$$

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.

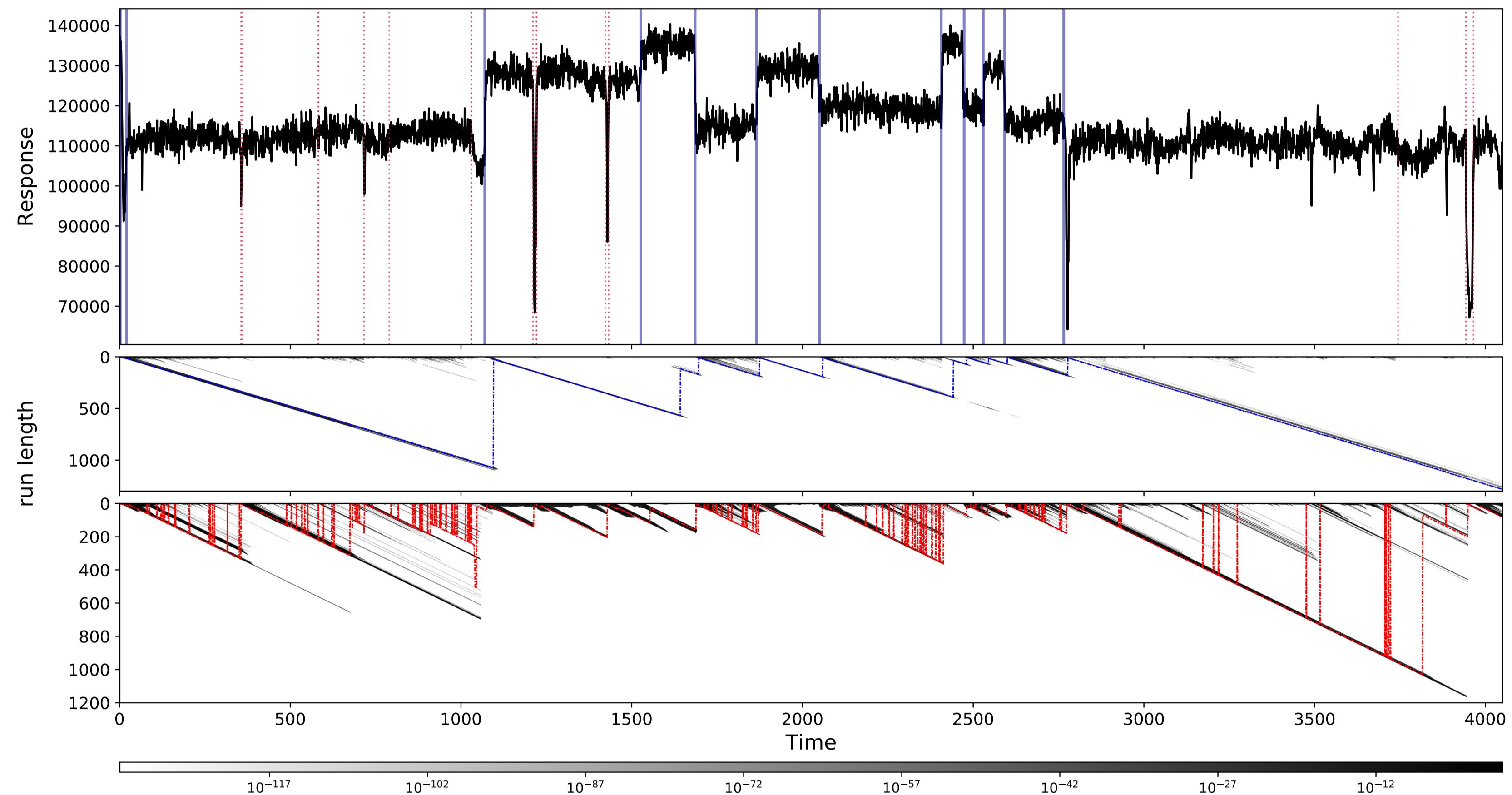


Figure 3: Top: well log data and changepoints found with robust BOCPD as solid lines. Additional changepoints found with standard BOCPD as dotted lines. Middle: Robust run-length posterior in grayscale, with emphasized maximum. Bottom: Standard run-length posterior in grayscale, with emphasized maximum.

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.

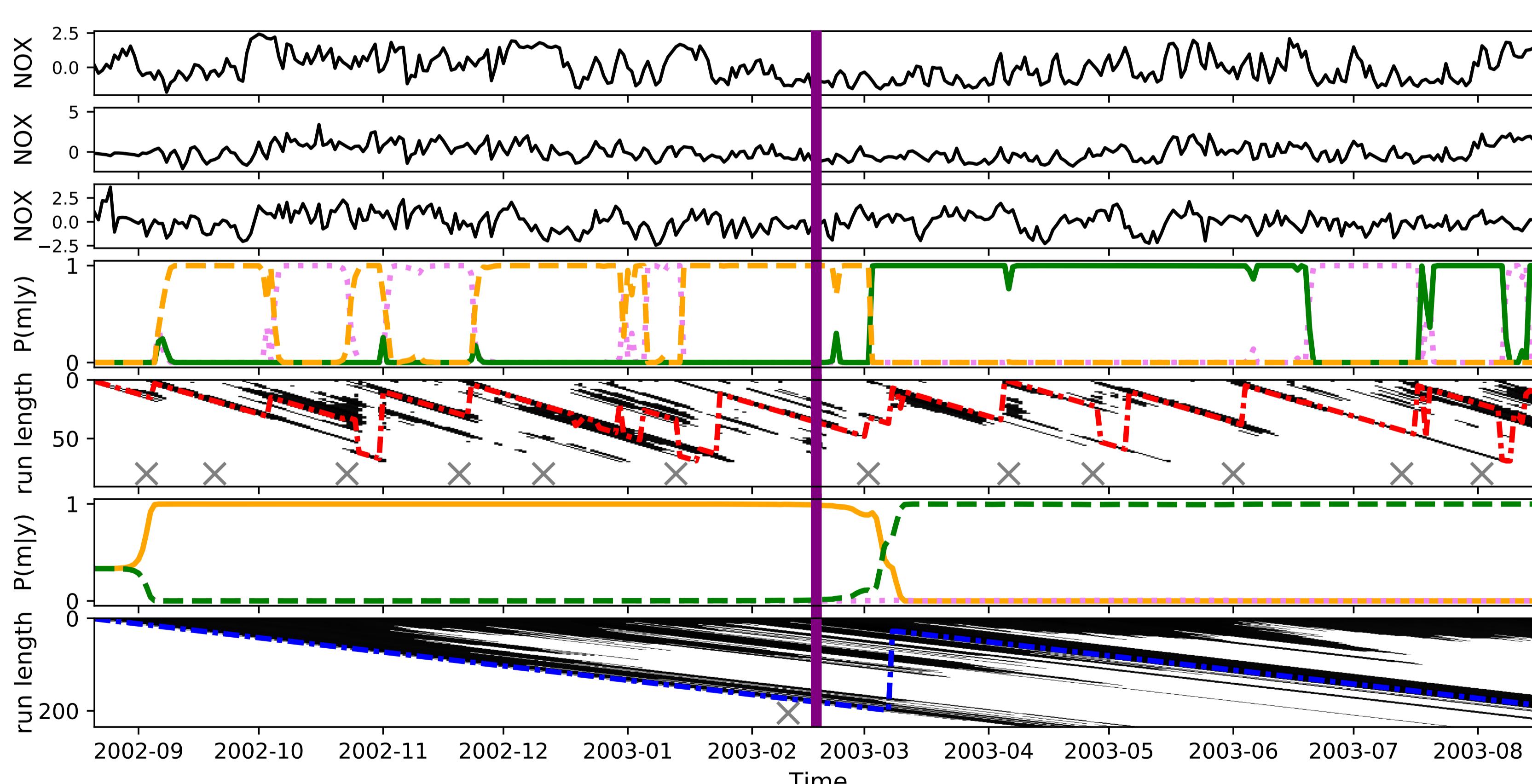


Figure 4: All Panels: Introduction of London's Congestion Charge as vertical line. Panels 1-3: Nitrogen Oxide measurements across London for 3/29 analyzed stations. Panels 4-5: On-line model posterior and run-length posteriors of standard BOCPD with detected changepoints marked as crosses (\times) and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

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

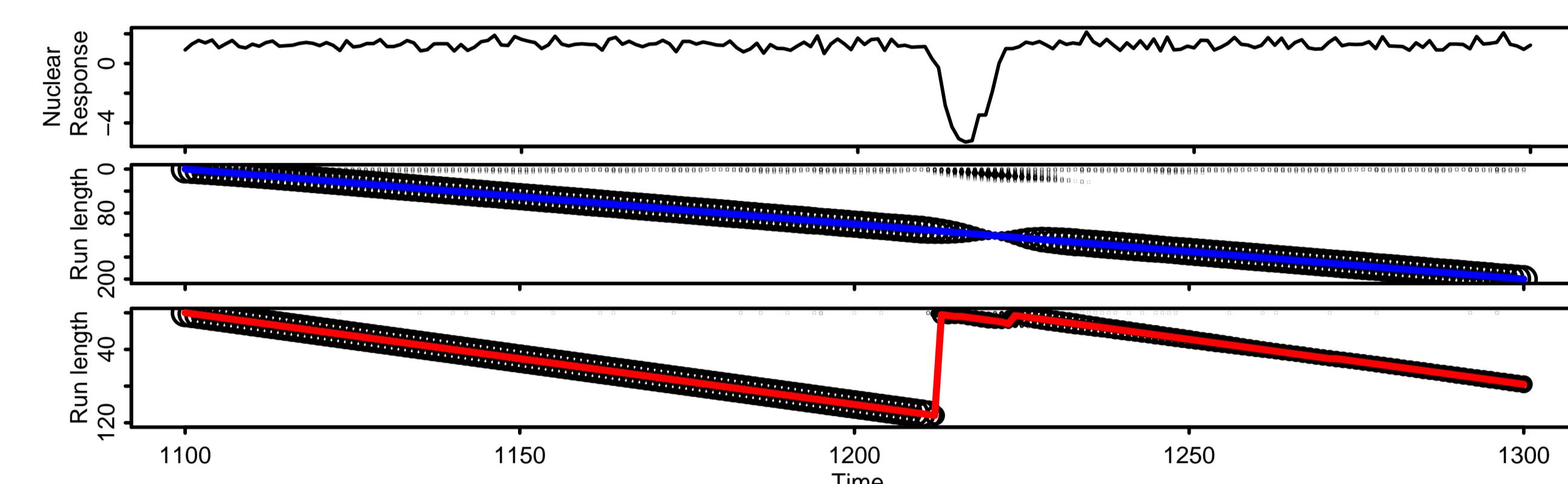


Figure 5: Top: 200 observations from the well log data. Middle: Gaussian β D run-length posterior. Bottom: Student's t_5 KLD run-length posterior.

THM. 2: EFFICIENT APPROXIMATION

One gets a closed form ELBO for the structural variational approximation

$$\hat{\pi}_m^{\beta_p}(\theta_m) = \underset{\pi_m^{\text{KLD}}(\theta_m)}{\operatorname{argmin}} \left\{ \text{KL} \left(\pi_m^{\text{KLD}}(\theta_m) \parallel \pi_m^{\beta_p}(\theta_m | \mathbf{y}_{(t-r_t):t}) \right) \right\}. \quad (9)$$

This means it is solvable with standard optimizers. $\hat{\pi}_m^{\beta_p}(\theta_m)$ is also attractive as it (I) is exact as $\beta_p \rightarrow 0$ 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:

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

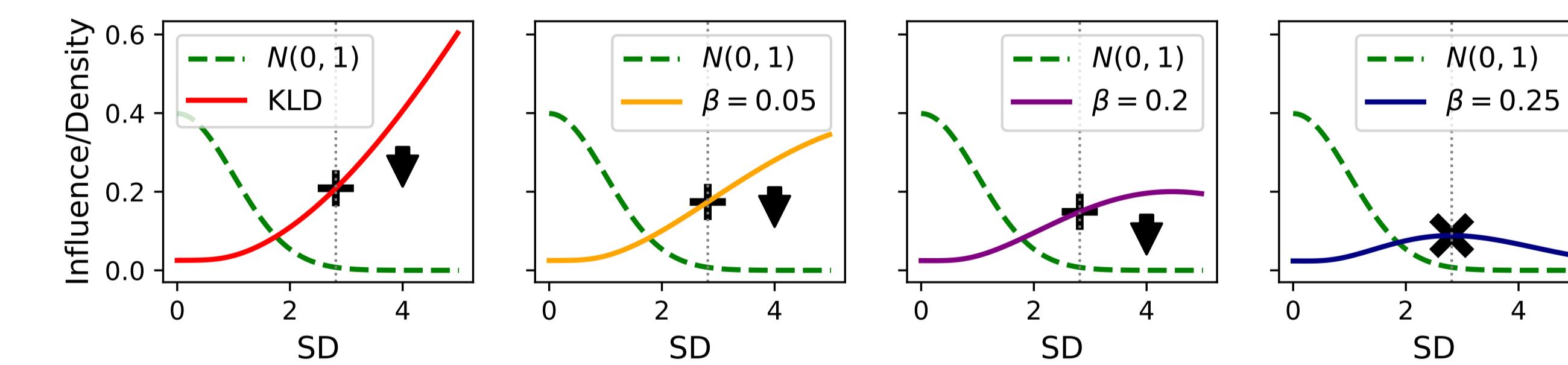


Figure 6: Initializing β by choosing a point of maximum influence

KEY REFERENCES

- Adams, R.P. & J.C. MacKay. 2007. 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

