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

Jeremias Knoblauch[1], Jack Jewson[1],
Theodoros Damoulas[2]

October 5, 2018

[1,2]University of Warwick, Department of Statistics
[2]University of Warwick, Department of Computer Science
[2]The Alan Turing Institute for Data Science and AI

Outline

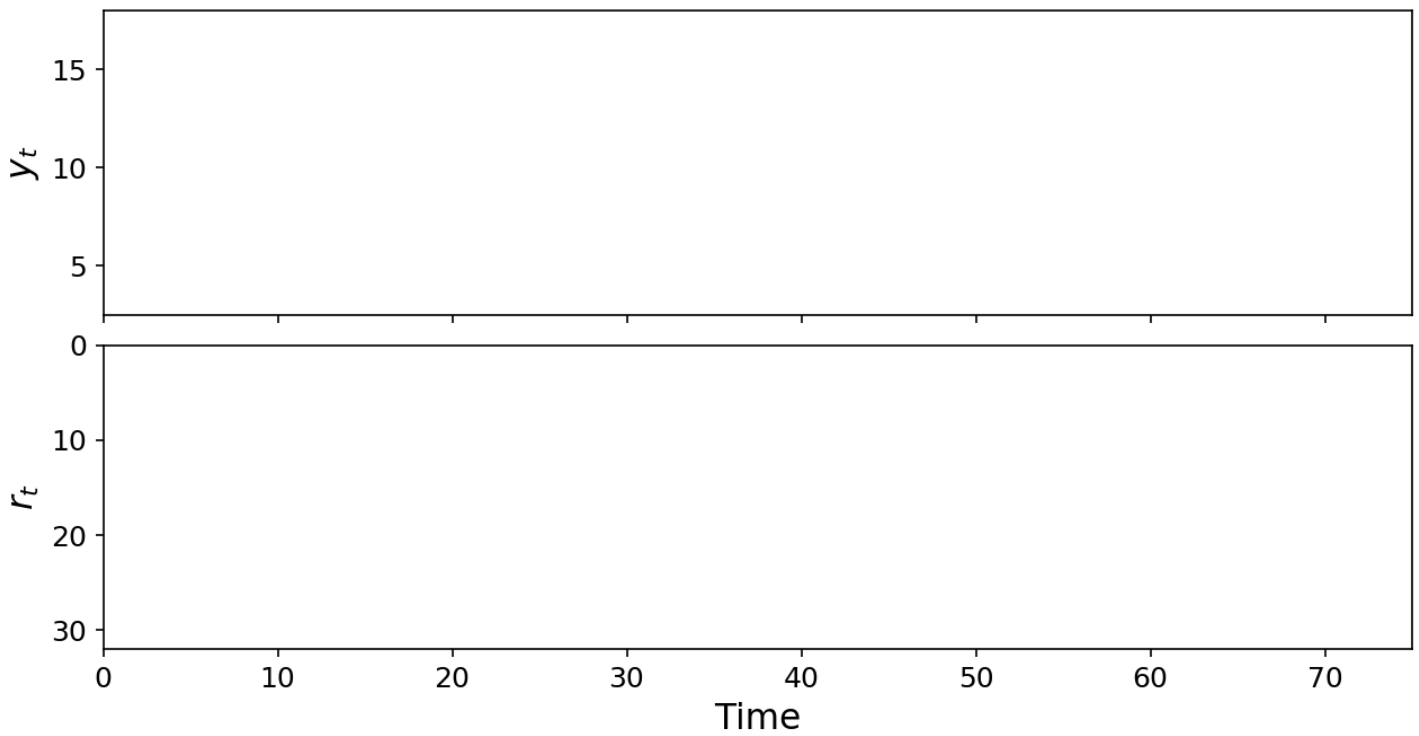
(1) BOCPD with Model Selection (Knoblauch and Damoulas, 2018):
- Intuition
- Computation & Inference
- Capabilities
- Limitations with outliers

(2) Robust CP detection with β-Divergences (Knoblauch et al., 2018):
- Intuition & Theory
- Computation & Inference
- New Capabilities

Standard Bayesian On-line Changepoint (CP) Detection

Idea due to Adams and MacKay (2007) and Fearnhead and Liu (2007):

(1) Define **Run-length at** $t = r_t \iff$ there was a CP at time $t - r_t$.
(2) **Inference on last CP** via $p(r_t|y_{1:t})$ rather than on *all* CPs
(3) Resulting complexity: $\mathcal{O}(t)$ **rather than** $\mathcal{O}(\prod_{i=1}^{t} i)$.



Recursion (with Model Selection) I/II

Idea: Combine multiple models (Fearnhead and Liu, 2007) with prediction (Adams and MacKay, 2007)
New Random Variable: m_t, the model at time t

$$r_t|r_{t-1} \sim H(r_t, r_{t-1}) \quad \text{[conditional CP prior]} \quad (1a)$$
$$m_t|m_{t-1}, r_t \sim q(m_t|m_{t-1}, r_t) \quad \text{[conditional model prior]} \quad (1b)$$
$$\boldsymbol{\theta}_m|m_t \sim \pi_{m_t}(\boldsymbol{\theta}_{m_t}) \quad \text{[parameter prior]} \quad (1c)$$
$$\boldsymbol{y}_t|m_t, \boldsymbol{\theta}_{m_t} \sim f_{m_t}(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t}) \quad \text{[observation density prior]} \quad (1d)$$

where $q(m_t|m_{t-1}, r_t) = \mathbb{1}_{\{r_t>0\}}\delta(m_{t-1}) + \mathbb{1}_{\{r_t=0\}}q(m_t)$.

Recursion:

$$p(\boldsymbol{y}_1, r_1 = 0, m_1) = q(m_1)\int_{\Theta_{m_1}} f_{m_1}(\boldsymbol{y}_1|\boldsymbol{\theta}_{m_1})\pi_{m_1}(\boldsymbol{\theta}_{m_1})d\boldsymbol{\theta}_{m_1} = q(m_1)f_{m_1}(\boldsymbol{y}_1|\boldsymbol{y}_0)$$

$$p(\boldsymbol{y}_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \Big\{ f_{m_t}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1})q(m_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) H(r_t, r_{t-1})p(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}$$

Recursion (with Model Selection) II/II

$$p(\boldsymbol{y}_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \Big\{ f_{m_t}(\boldsymbol{y}_t | \boldsymbol{y}_{1:(t-1)}, r_{t-1}) q(m_t | \boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \\ H(r_t, r_{t-1}) p(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}$$

Inference:

(1) Evidence: $p(\boldsymbol{y}_{1:t}) = \sum_{r_t, m_t} p(\boldsymbol{y}_{1:t}, r_t, m_t)$

(2) run-length & model posterior: $p(r_t, m_t | \boldsymbol{y}_{1:t}) = p(\boldsymbol{y}_{1:t}, r_t, m_t) / p(\boldsymbol{y}_{1:t})$

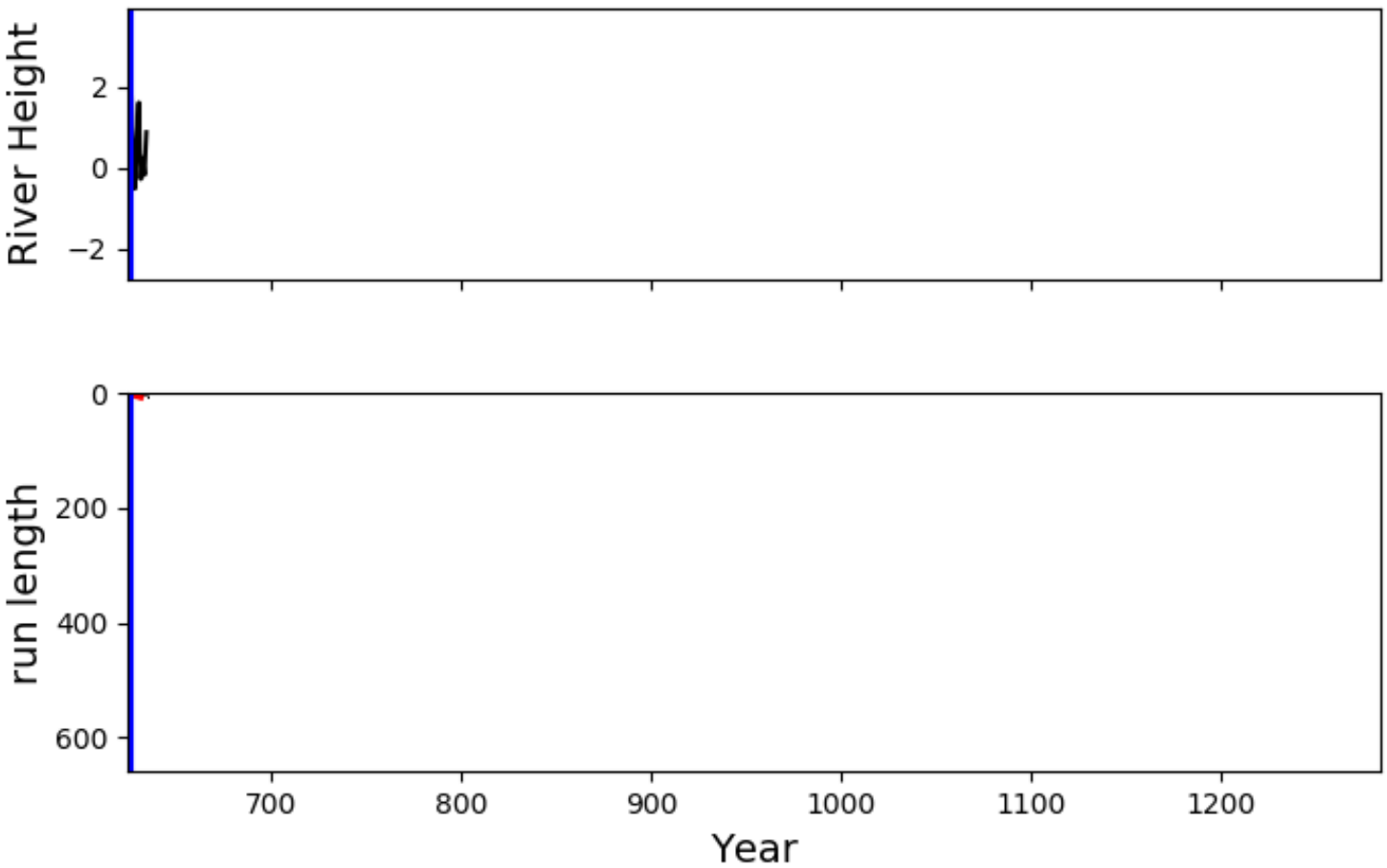
(3) Prediction: $p(\boldsymbol{y}_{t+1} | \boldsymbol{y}_{1:t}) = \sum_{r_t, m_t} f_{m_t}(\boldsymbol{y}_{t+1} | \boldsymbol{y}_{1:t}, r_t) p(r_t, m_t | \boldsymbol{y}_{1:t})$

(4) Run-length marginal posterior: $p(r_t | \boldsymbol{y}_{1:t}) = \sum_{m_t} p(r_t, m_t | \boldsymbol{y}_{1:t})$

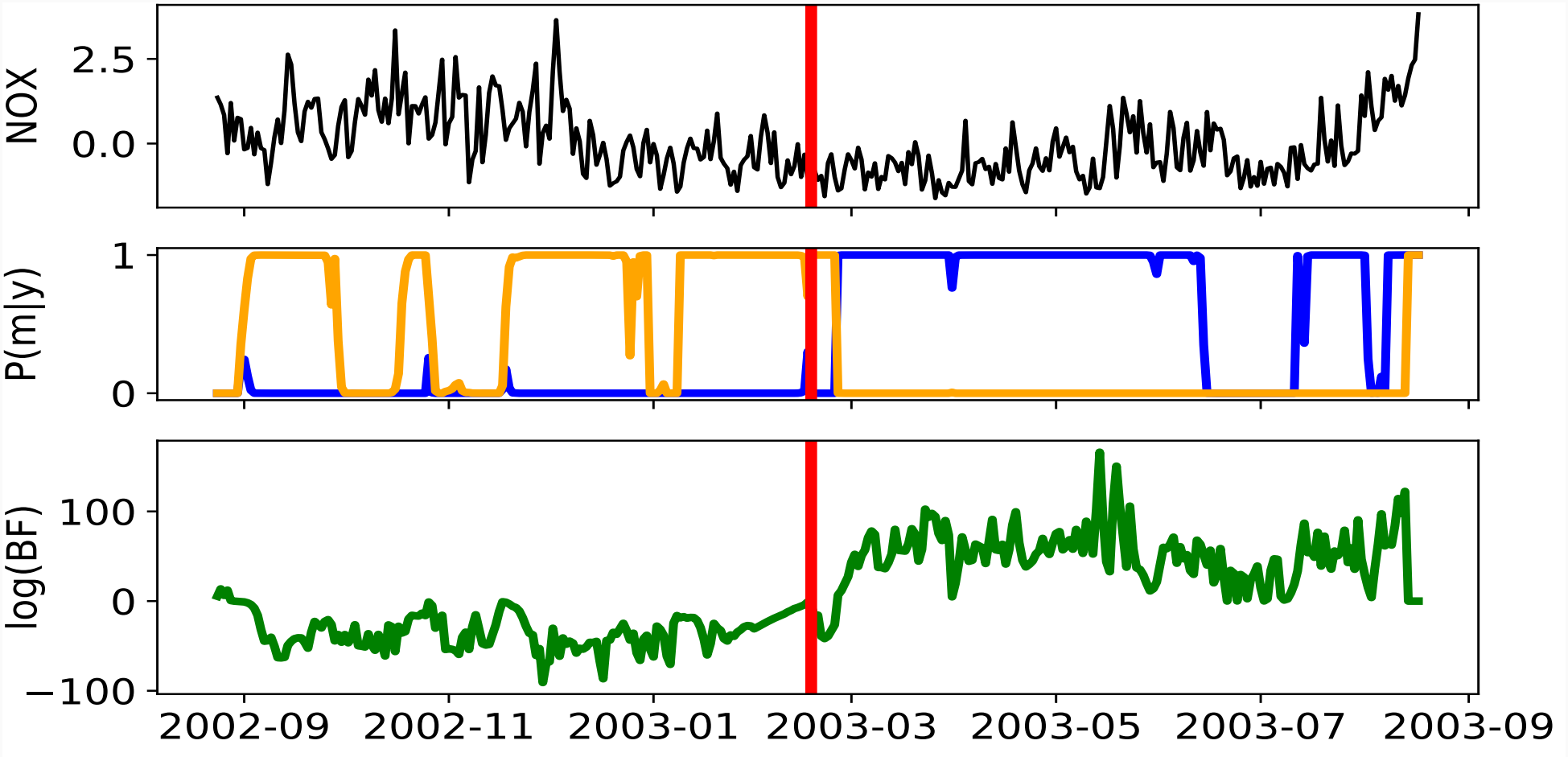
(5) Model marginal posterior: $p(m_t | \boldsymbol{y}_{1:t}) = \sum_{r_t} p(r_t, m_t | \boldsymbol{y}_{1:t})$.

(6) MAP segmentation:
$MAP_t = \max_{r,t} \{ MAP_{t-r-1} \cdot p(r_t = r, m_t = m | \boldsymbol{y}_{1:t}) \}$

CP detection [e.g., Nile data]



On-line Model Selection on shifting multivariate dynamics



Panel 1: NOX levels in London with **congestion charge introduction**

Panel 2: Model posteriors for the two VAR models

Panel 3: Corresponding log Bayes Factors

Limitation: BOCPD is not robust to outliers/misspecification

Why is it non-robust?

- On-line processing
- Moderate/high dimensions for $\boldsymbol{y}_t$

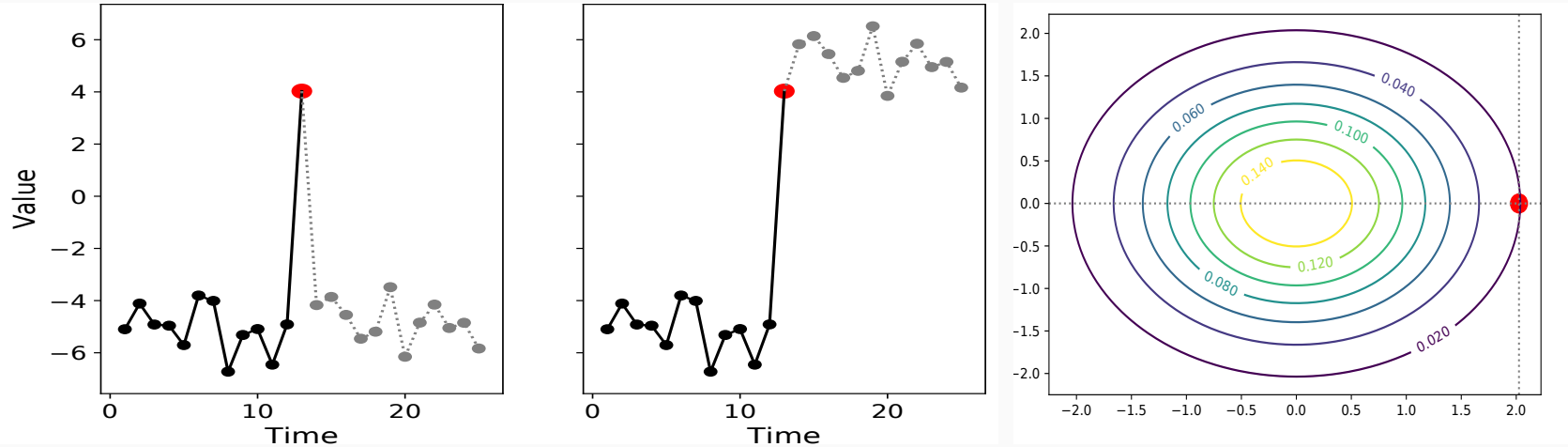
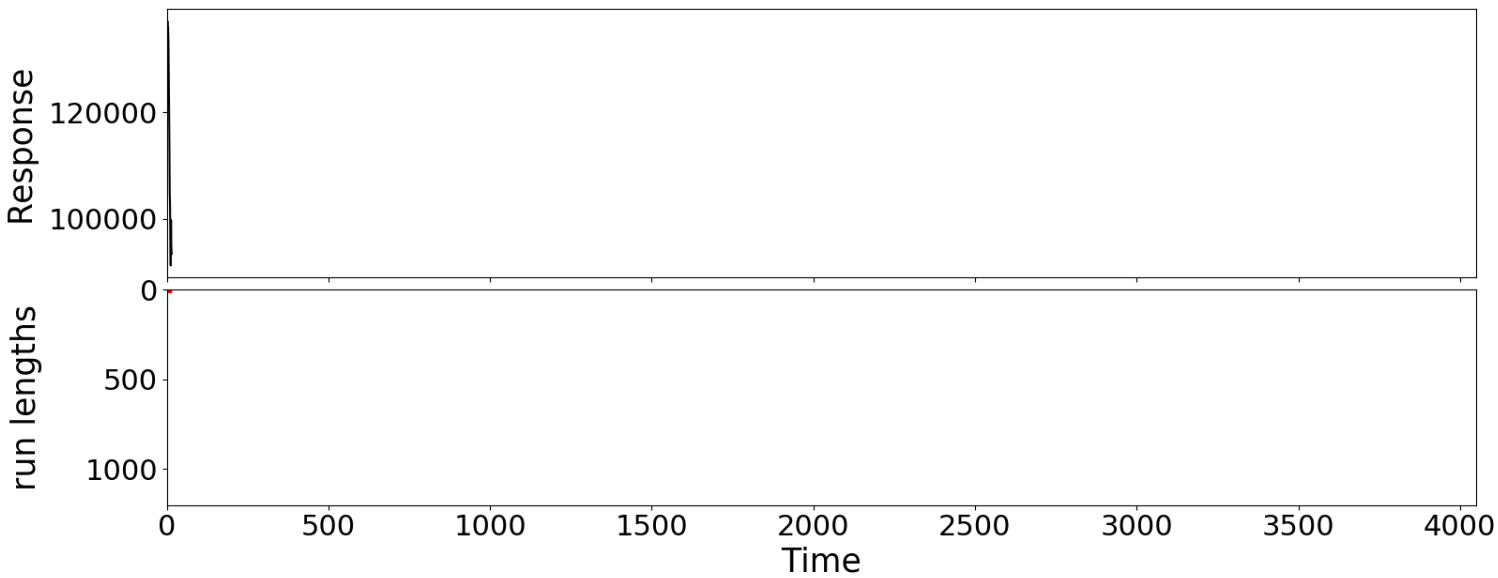


Figure 1 – Left, Center: Price for on-line processing is that outliers are confused with changepoings. **Right:** Multivariate densities become very small even if outliers occur only in a single dimension.

Illustration: BOCPD and the well-log data



Robust CP detection in high-dimensional data streams

Five Autoregressive processes with two CPs

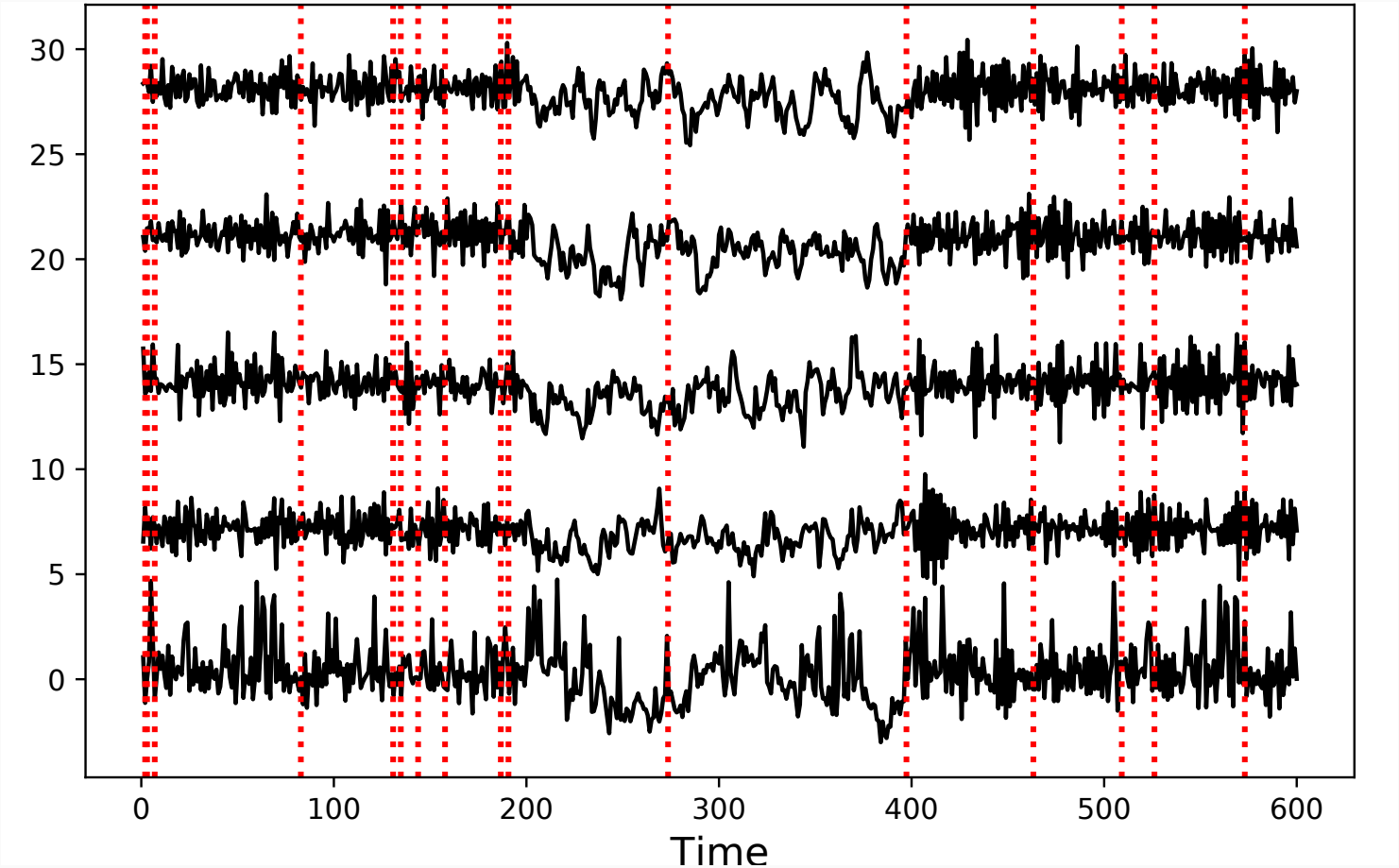


Figure 2 – Maximum A Posteriori (MAP) CPs of standard BOCPD shown as dashed vertical lines. True CPs at $t = 200, 400$.

What are outliers (= model misspecification) really?

	M-closed world	**M-open world**
Model	There exists θ_0 s.t. $Y \sim f(\cdot;\theta_0)$	f is a decent description of Y's sampling distribution
Misspecification	Such a θ_0 does *not* exist	*All models are wrong but some are useful*
Optimal Divergence	Kullback-Leibler	*no* optimal divergence

 Outliers represent a form of misspecification

 if f is misspecified w.r.t. Y, we *cannot* learn θ_0

 If outliers are severe, inference in the M-closed world breaks down

 We address this using the M-open paradigm

$\Longrightarrow$ See Jewson et al. (2018) and Bissiri et al. (2016)

The M-open paradigm: General Bayesian Updating I/II

Observation 1: Making decisions like a Bayesian means finding

$$\theta^* = \arg\min_{\theta} \int_{\mathcal{Y}} \ell(\theta, y) dG, \tag{3}$$

where $Y \sim G$ and ℓ is a loss coupling y to θ.

Observation 2: Bissiri et al. (2016) solve original Bayesian problem, updating beliefs about θ only a loss ℓ. *No* model for G, i.e. θ could be the median, e.g. Approach yields the generalized posterior

$$\pi(\theta|\mathbf{y}_{1:T}) \propto \pi(\theta) \exp\left(-\sum_{t=1}^{T} \ell(\theta, y_t)\right) \tag{4}$$

Observation 3: In the M-closed world/standard Bayesian inference, model family f known. $\implies$ now, $\ell(\theta, y) = \log(f(y|\theta))$; is standard Bayes rule and minimizes Kullback-Leibler divergence between f and G

The M-open paradigm: General Bayesian Updating II/II

Observation 4: Problem: Standard Bayesian Inference/Kullback-Leibler divergence is **not robust** [see picture next slide]

Observation 5: Every divergence D can be expressed in terms of some loss ℓ_D. $\Longrightarrow$ if we want to minimize divergence D between f and G (Jewson et al., 2018):

$$\pi^D(\theta|\mathbf{y}_{1:T}) \propto \pi^D(\theta) \exp\left(-\sum_{t=1}^{T} \ell_D(x_t, f(\cdot;\theta))\right). \tag{5}$$

Observation 6: Divergences of the α, β, γ families are especially suited to deal with outliers.

So we use General Bayesian Inference (GBI) to robustify BOCPD

Solution: Robustness via Generalized Bayesian Inference

Illustration of β-Divergence (β-D) robustness via influence functions

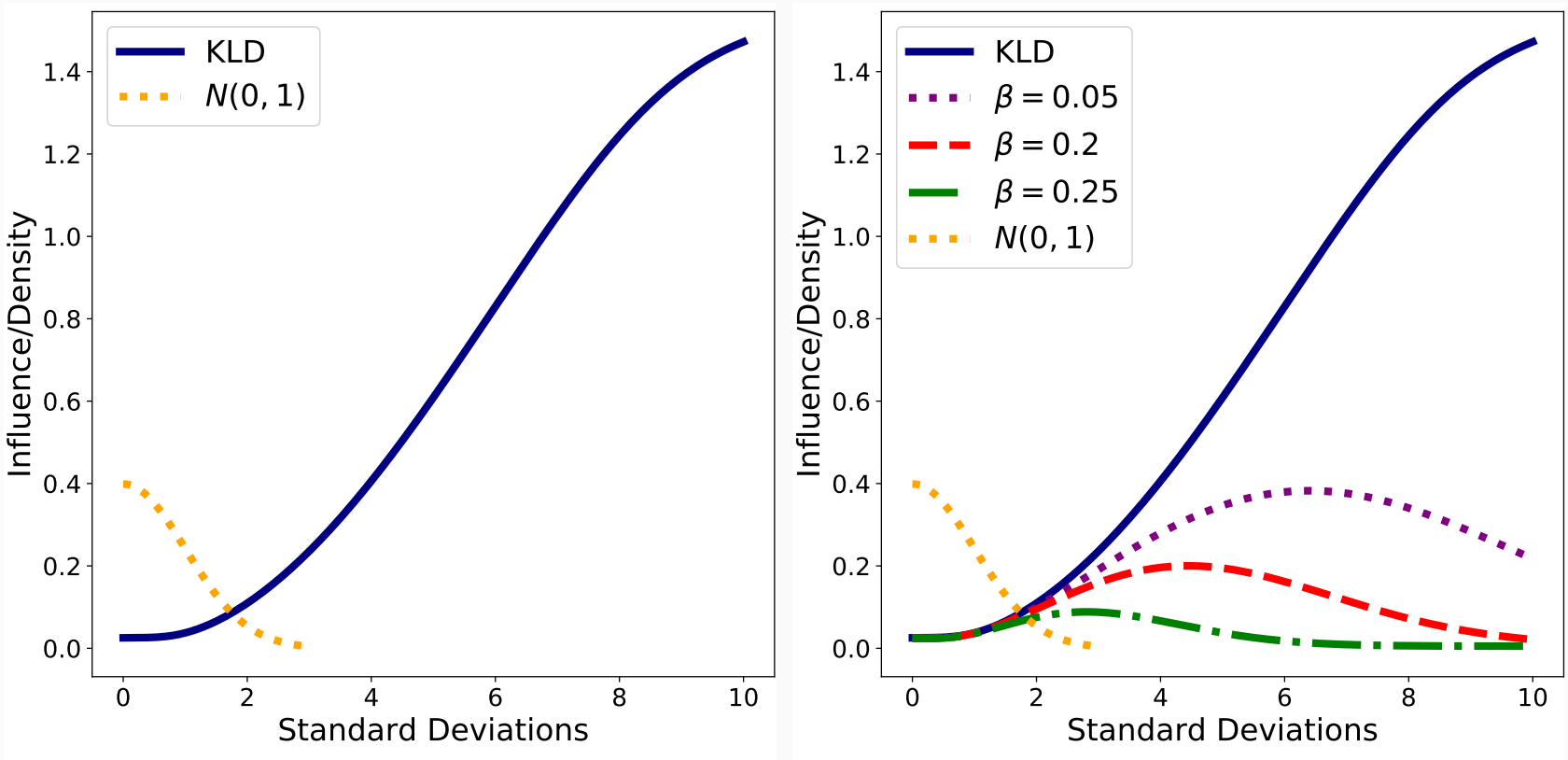


Figure 3 – Left: standard Bayes influence (Kullback-Leibler Divergence (KLD)) and standard normal density. **Right**: Robust β-Divergence (β-D) family.

Robust Recursion for BOCPD

We propose β-D-based Generalized Bayesian Inference (GBI) to make BOCPD **doubly robust**: For the inference on $\boldsymbol{\theta}_m$ *and* on (r_t, m_t).

Parameter Layer robustified via β_{p}:

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

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

$$f_m(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}) = \int_{\Theta_m} f_m(\boldsymbol{y}_t|\boldsymbol{\theta}_m)\pi_m^{\beta}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_{t-1}):(t-1)}) d\boldsymbol{\theta}_m$$

Run-length and Model Layer robustified via β_{rlm}:

$$\widetilde{f}_m(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}) = e^{-\left(\frac{1}{\beta_{\mathsf{rlm}}} f_m(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1})^{\beta_{\mathsf{rlm}}} - \frac{1}{1+\beta_{\mathsf{rlm}}} \int_{\mathcal{Y}} f_m(\boldsymbol{z}|\boldsymbol{y}_{1:(t-1)}, r_{t-1})^{1+\beta_{\mathsf{rlm}}} d\boldsymbol{z}\right)}$$

Inference/Recursion:

$$p^{\beta}(\boldsymbol{y}_{1:t}, r_t, m_t) \propto \sum_{m_{t-1}, r_{t-1}} \Big\{ \widetilde{f}_{m_t}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}) q(m_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) H(r_t, r_{t-1}) p^{\beta}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}$$

Robust & Fast Computation for BOCPD

β-D posterior not scalable $\Longrightarrow$ Structural Variational Inference

Observation I: As $\beta \to 0$, β-D $\to$ KLD! $\Longrightarrow \pi_m^{\mathrm{KLD}} \approx \pi_m^{\beta}$ for small β!

Observation II: In fact, we prove that for most conjugate exponential family models, we get a closed-form ELBO objective approximating

$$\widehat{\pi}_m^{\beta_{\mathrm{p}}}(\boldsymbol{\theta}_m) = \underset{\pi_m^{\mathrm{KLD}}(\boldsymbol{\theta}_m)}{\operatorname{argmin}} \left\{ \mathrm{KL}\left(\pi_m^{\mathrm{KLD}}(\boldsymbol{\theta}_m) \middle\| \pi_m^{\beta_{\mathrm{p}}}(\boldsymbol{\theta}_m | \boldsymbol{y}_{(t-r_t):t}) \right) \right\}. \qquad (6)$$

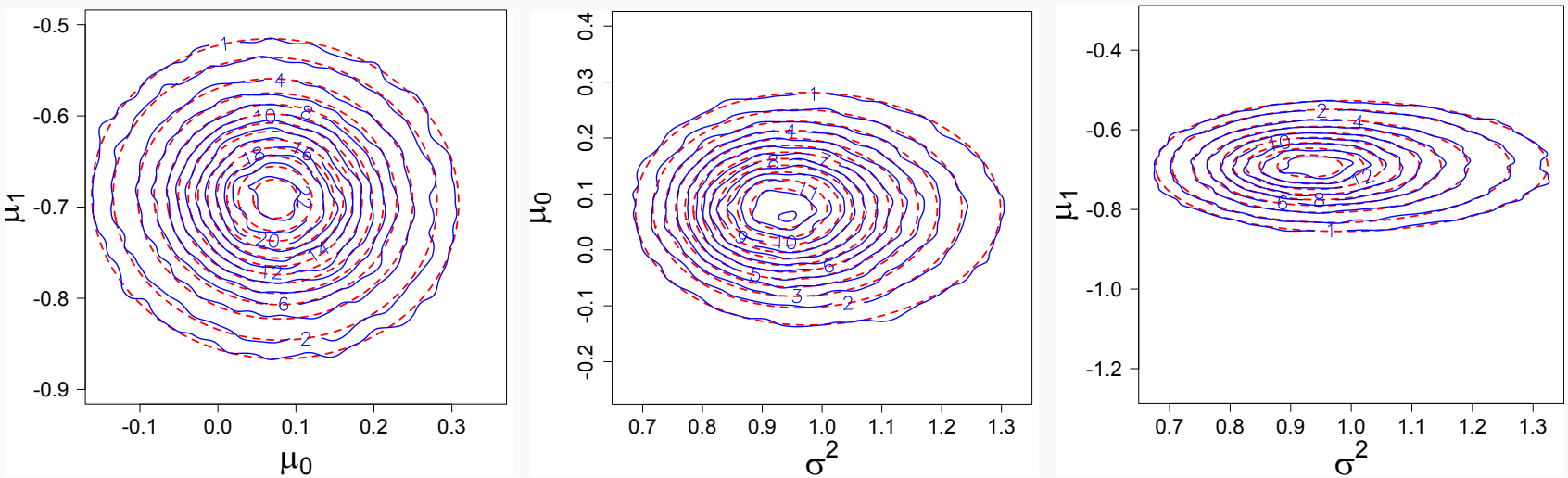
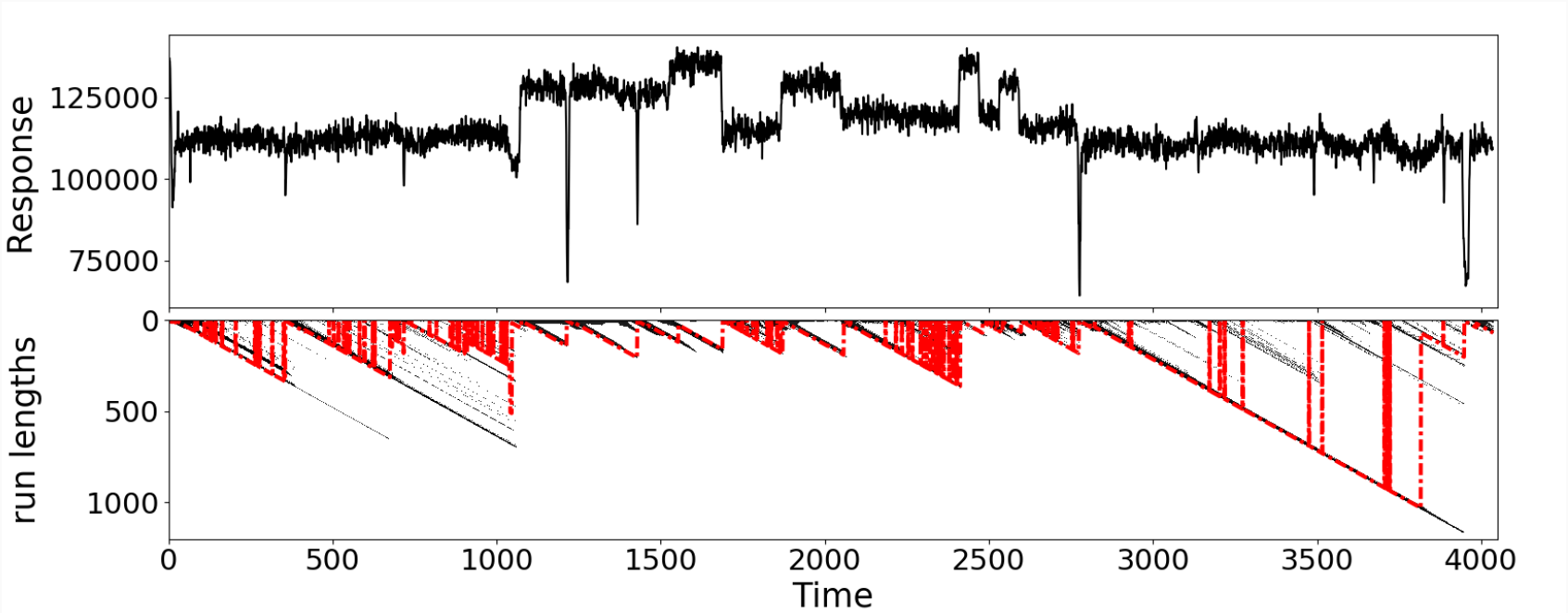


Figure 4 – Contour plots of bivariate marginals of approximation $\widehat{\pi}_m^{\beta_{\mathrm{p}}}(\boldsymbol{\theta}_m)$ (dashed) and the target $\pi_m^{\beta_{\mathrm{p}}}(\boldsymbol{\theta}_m | \boldsymbol{y}_{(t-r_t):t})$ (solid) estimated from 95,000 Hamiltonian Monte Carlo samples for BLR ($d = 1$, two regressors, $\beta_{\mathrm{p}} = 0.25$).

Remember this?



Robust CP detection in outlier-prone data streams

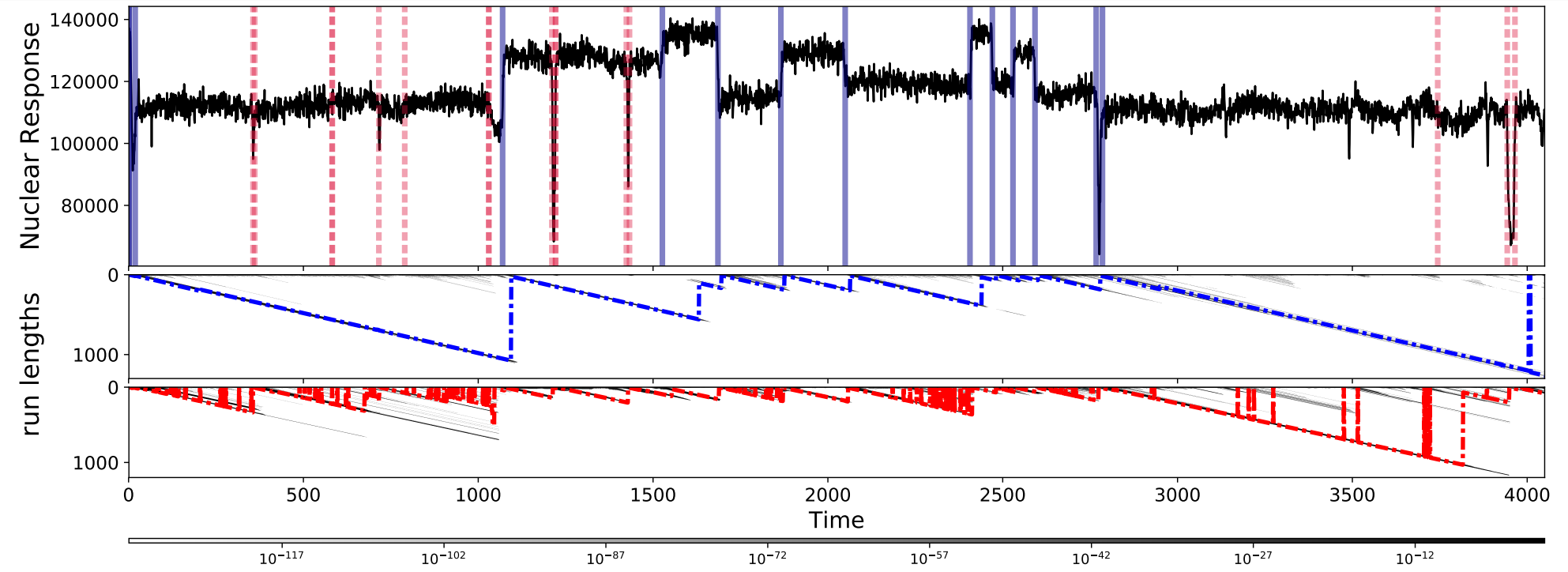


Figure 5 – Robust segmentation and run-length distribution and **additionally found CPs with non-robust run-length distribution**

[FDR: $> 99\% \Longrightarrow 8\%$ and reduction in MSE (MAE) by 10% (6%)]

Remember this?

Five Autoregressive processes with two CPs

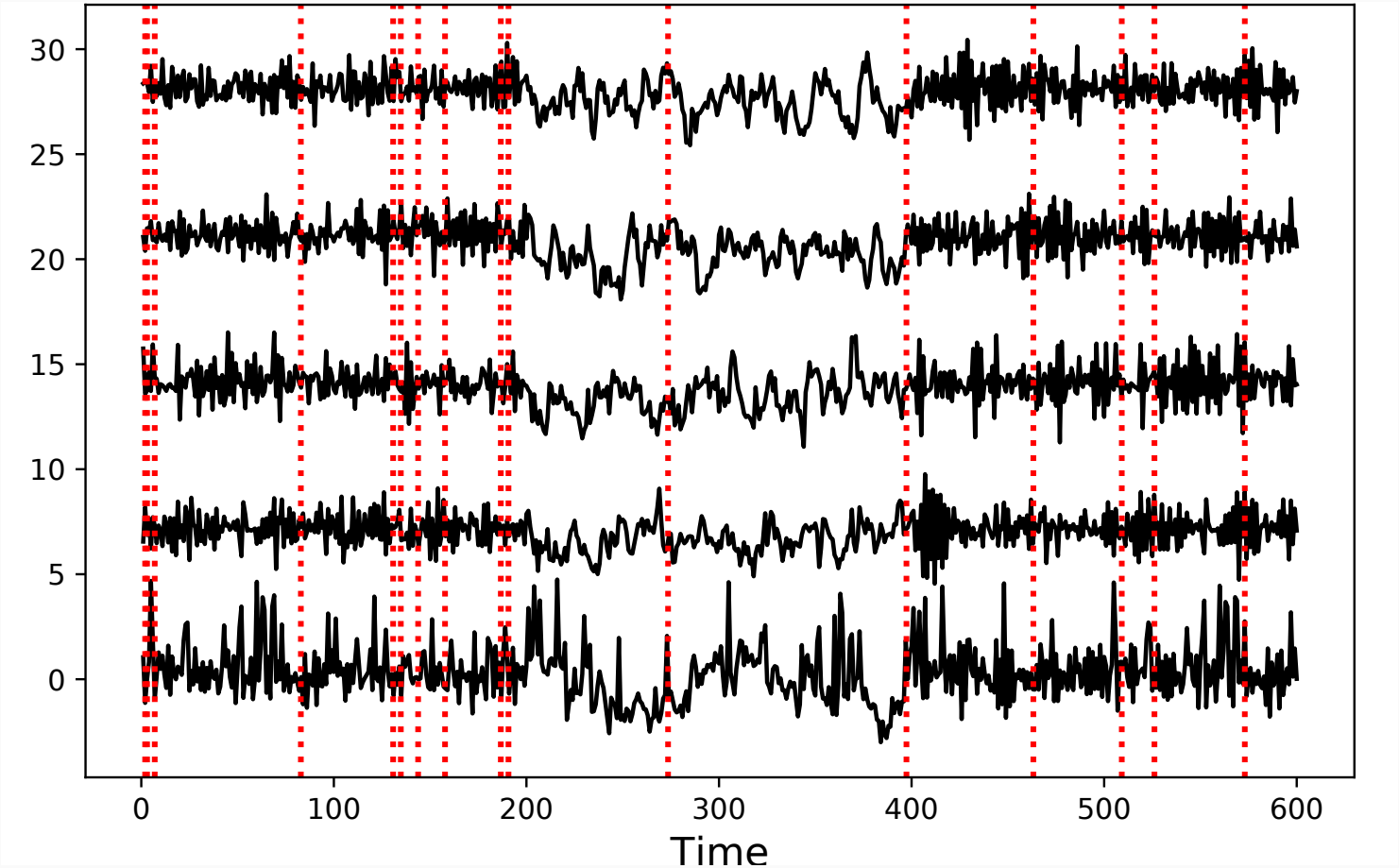


Figure 6 – Maximum A Posteriori (MAP) CPs of standard BOCPD shown as dashed vertical lines. True CPs at $t = 200, 400$.

Robust CP detection in high-dimensional data streams II/II

Five Autoregressive processes with two CPs

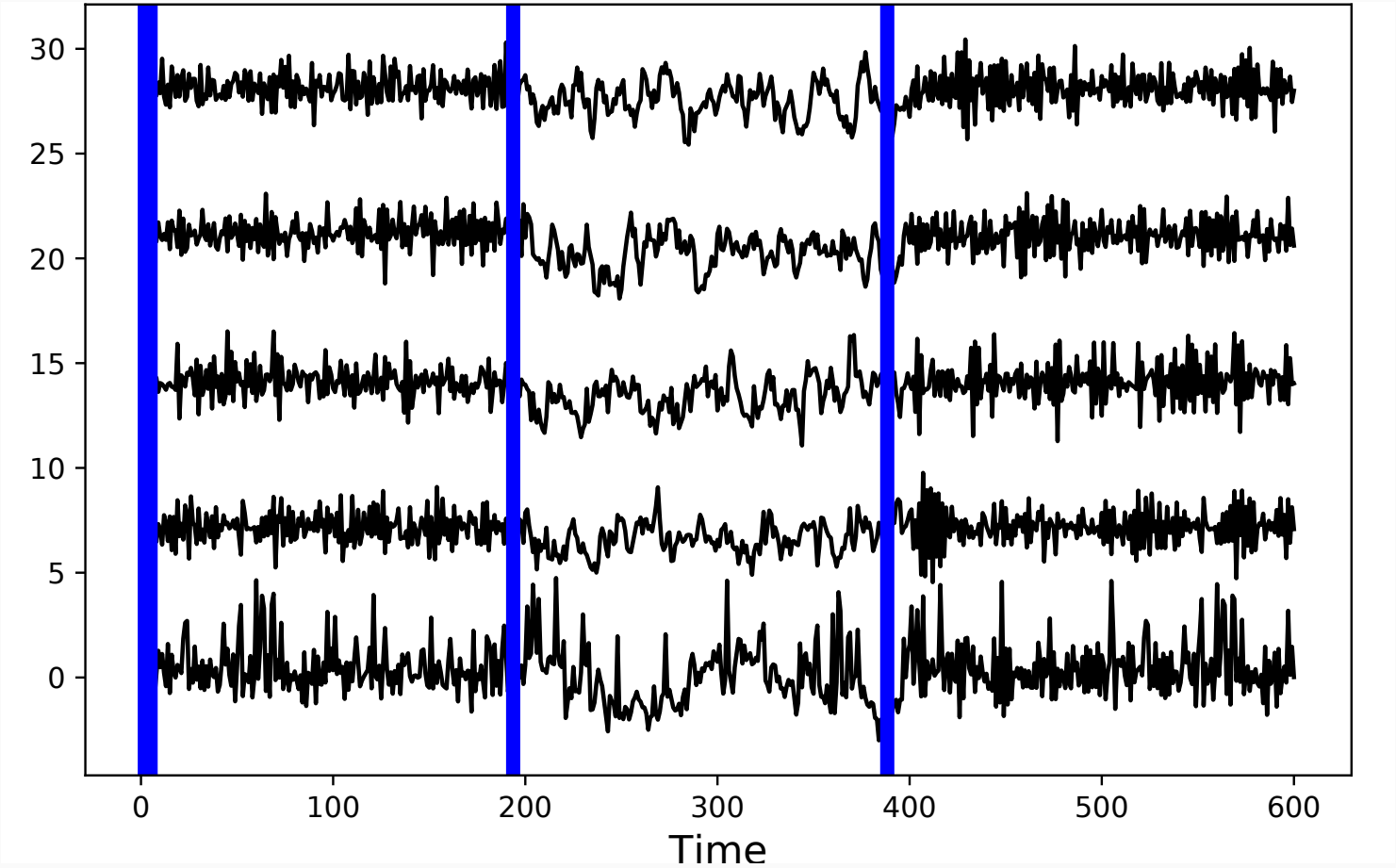
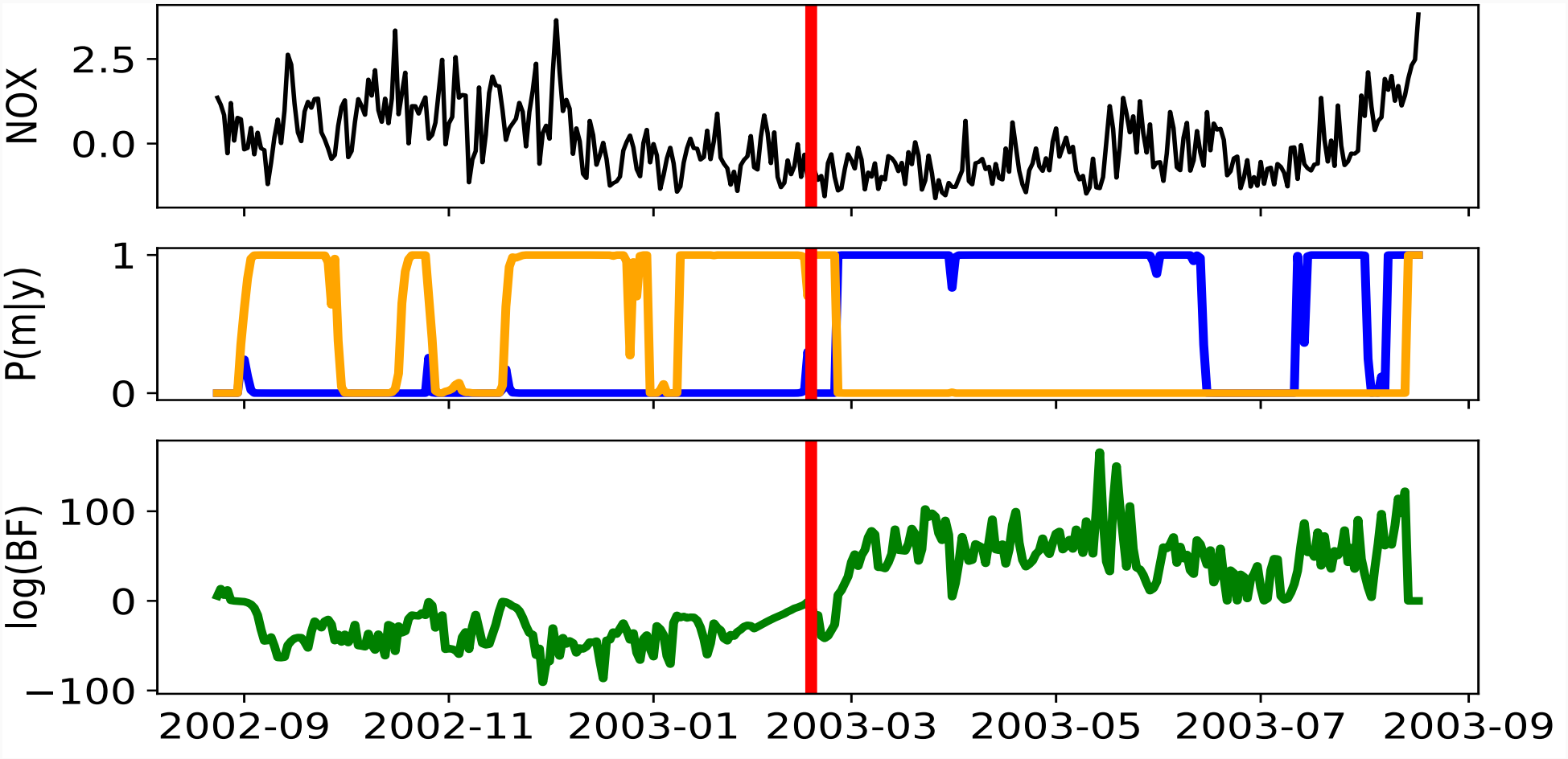


Figure 7 – Maximum A Posteriori (MAP) CPs of robust BOCPD shown as solid vertical lines. True CPs at $t = 200, 400$.

Remember this?



Panel 1: NOX levels in London with **congestion charge introduction**

Panel 2: Model posteriors for the two VAR models

Panel 3: Corresponding log Bayes Factors

Robust Model Selection on shifting multivariate dynamics

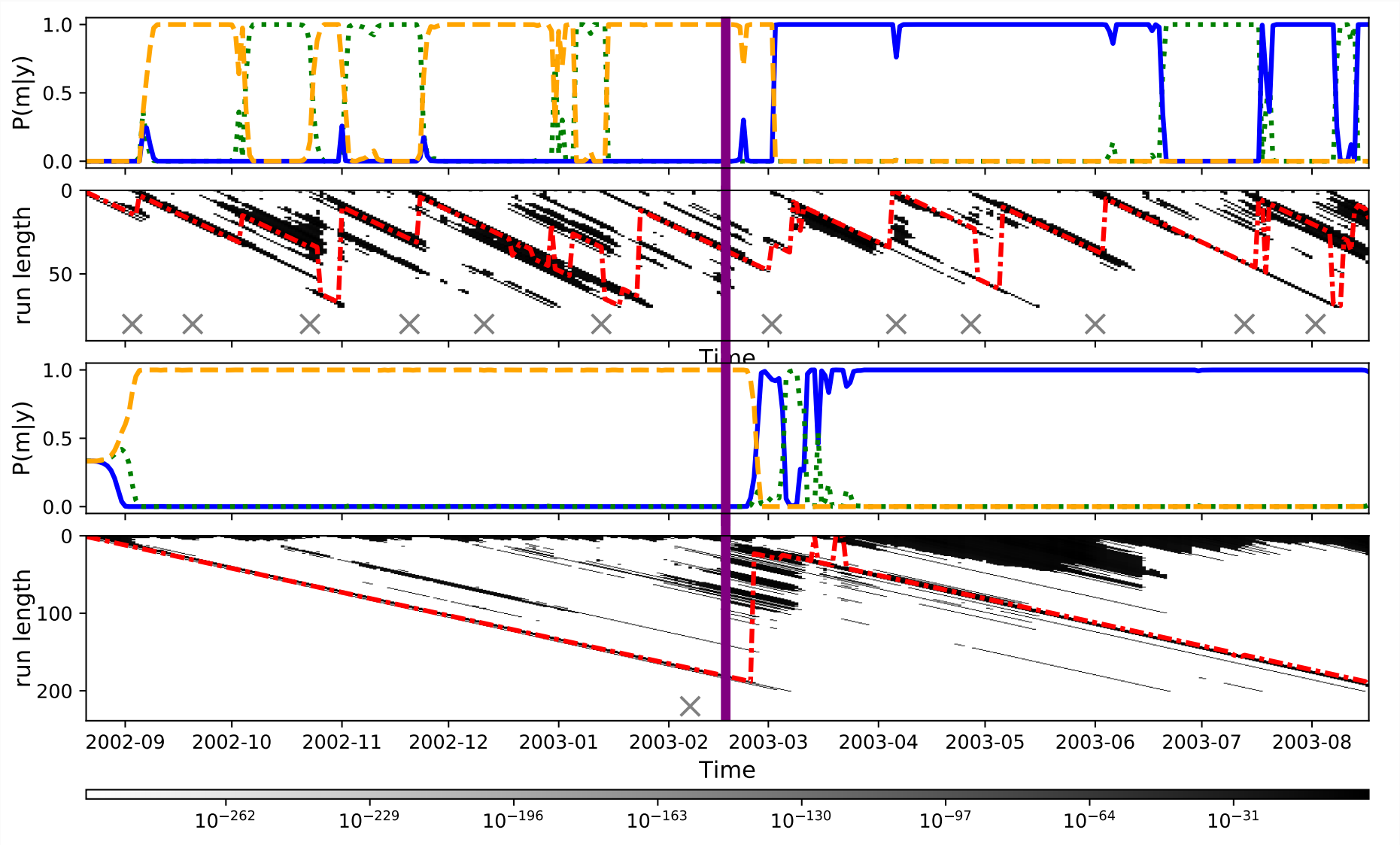


Figure 8 – Top & bottom two panels: standard & robust BOCPD.

Main References

Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

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

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

Jewson, J., Smith, J. Q., and Holmes, C. (2018). Principled bayesian minimum divergence inference. *arXiv preprint arXiv:1802.09411*.

Knoblauch, J. and Damoulas, T. (2018). Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-18)*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2018). Doubly robust bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NIPS)*. to appear.

Q: How to initialize β (= level of robustness)?

Choice of $\beta \Longrightarrow$ Initialization.

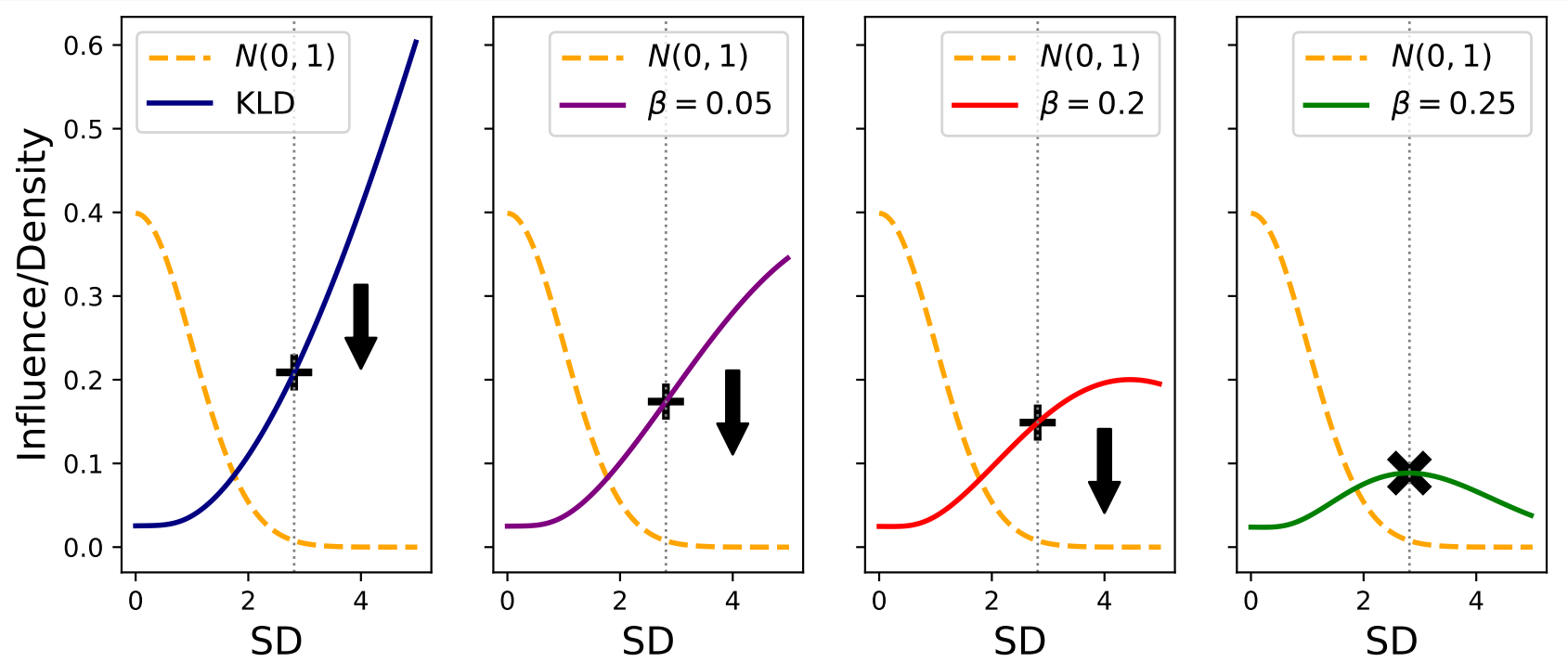


Figure 9 – Initialization procedure visualized for a (standard) normal prior on $y_t \in \mathbb{R}$ and if we want to have maximum influence of an observation 2.75 standard deviation units from the expected value under the current belief.

Q: How to refine β (= level of robustness)?

Bad initialization of $\beta \Longrightarrow$ On-line optimization using SGD

Idea: For a predictive loss function L, and prediction $\widehat{\boldsymbol{y}}_t(\beta)$, apply SGD to $L(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t(\beta)$ w.r.t. β

[For $p^{\beta_{\text{rlm}}}(\boldsymbol{y}_{1:t}, r_t, m_t)$: closed form gradients available; For $\widehat{\pi}_m^{\beta_{\text{p}}}(\boldsymbol{\theta}_m)$: Numerical gradient approximations used]

One can then minimize L on-line via

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - \eta \cdot \begin{bmatrix} \nabla_{\beta_{\text{rlm},t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right) \\ \nabla_{\beta_{\text{p},t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right)) \end{bmatrix} \tag{7}$$