



WARWICK
THE UNIVERSITY OF WARWICK

OxWaSP

The
Alan Turing
Institute



Lloyd's Register
Foundation

Robust Spatio-Temporal Bayesian On-line Changepoint Detection and Model Selection

Jeremias Knoblauch¹, Theodoros Damoulas²,

Jack Jewson¹

July 6, 2018



^{1,2}University of Warwick, Department of Statistics

²University of Warwick, Department of Computer Science

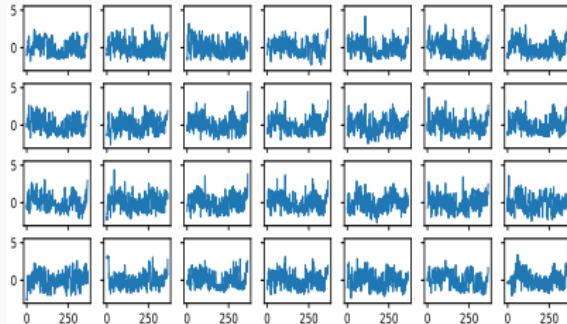
²The Alan Turing Institute for Data Science and AI

Outline

- (1) **Motivation** for BOCPD
- (2) **Standard** BOCPD: Overview,  Limitations
- (3)  **Extension I:** Multivariate/Spatio -Temporal Models
- (4)  **Extension II:** Model Selection
- (5)  **Extension III:** Robustification

[**Standard** BOCPD presented as in Adams and MacKay (2007);
Extensions I and **II** in Knoblauch and Damoulas (2018);
Extension III in Knoblauch et al. (2018)]

Motivation: Features of Air Pollution in London



- (1) **Spatio-temporal** data stream
- (2) Want to do principled probabilistic (= **Bayesian**) forecasting
- (3) Observation frequency high \Rightarrow **on-line** treatment essential
- (4) Abrupt changes (congestion charge) \Rightarrow **Changepoint Detection**

\Rightarrow **Spatio-temporal Bayesian On-line Changepoint Detection**

Extension I

Adams and MacKay (2007)

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** $\mathcal{O}(\prod_{i=1}^t i)$.

Standard BOCPD: Probabilistic model & Inference

$$r_t | r_{t-1} \sim H(r_t, r_{t-1}) \quad [\text{conditional CP prior}] \quad (1a)$$

$$\theta \sim \pi(\theta) \quad [\text{parameter prior}] \quad (1b)$$

$$y_t | \theta \sim f(y_t | \theta) \quad [\text{observation density prior}] \quad (1c)$$

implicitly, a requirement of (f, π) is that the posterior predictives

$$f(y_t | y_{1:(t-1)}, r_{t-1}) = \int_{\Theta} f(y_t | \theta) \pi(\theta | y_{(t-r_{t-1}): (t-1)}) d\theta \quad (2)$$

are efficiently computable. Inference then proceeds via the recursion

$$p(y_1, r_1 = 0) = \int_{\Theta} f(y_1 | \theta) \pi(\theta) d\theta = f(y_1 | y_0), \quad (3)$$

$$p(y_{1:t}, r_t) = \sum_{r_{t-1}} \left\{ f(y_t | y_{1:(t-1)}, r_{t-1}) H(r_t, r_{t-1}) p(y_{1:(t-1)}, r_{t-1}) \right\}. \quad (4)$$

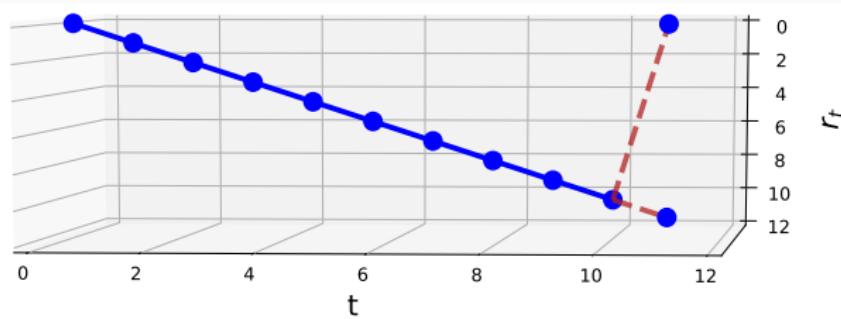
Observations:

- (1) Assumes that the same model (f, π) holds in each segment
- (2) CPs are shifts in the parameterization θ of that model

Standard BOCPD: Inference

$$p(\mathbf{y}_1, r_1 = 0) = \int_{\Theta} f(\mathbf{y}_1 | \theta) \pi(\theta) d\theta = f(\mathbf{y}_1 | \mathbf{y}_0),$$

$$p(\mathbf{y}_{1:t}, r_t) = \sum_{r_{t-1}} \left\{ f(\mathbf{y}_t | \mathbf{y}_{1:(t-1)}, r_{t-1}) H(r_t, r_{t-1}) p(\mathbf{y}_{1:(t-1)}, r_{t-1}) \right\}.$$



Inference:

- (1) **Evidence:** $p(\mathbf{y}_{1:t}) = \sum_{r_t} p(\mathbf{y}_{1:t}, r_t)$
- (2) **CP (run-length) posterior:** $p(r_t | \mathbf{y}_{1:t}) = p(\mathbf{y}_{1:t}, r_t) / p(\mathbf{y}_{1:t})$.
- (3) **Prediction:** $p(\mathbf{y}_{t+1} | \mathbf{y}_{1:t}) = \sum_{r_t} f(\mathbf{y}_{t+1} | \mathbf{y}_{1:t}, r_t) p(r_t | \mathbf{y}_{1:t})$
- (4) **MAP segmentation:** $MAP_t = \max_r \{ MAP_{t-r-1} p(r_t = r | \mathbf{y}_{1:t}) \}$

Standard BOCPD: Illustration using AR(1) on Nile data



Standard BOCPD: Limitations

Limitations:

- So far: inherently **univariate** method
- Each segment described by the **same model** $m = (f, \pi)$
- non-robust** to outliers and misspecification

Contributions:

- Construct inherently **multivariate** models m
- Allow **multiple models** $\{(f_1, \pi_1), \dots, (f_K, \pi_K)\}$ for the segments
- Update your posteriors with **robust** divergences



Extension I: Multivariate/Spatio-Temporal models

Ingredient 1: Standard Bayesian VAR model

$$\sigma^2 \sim \text{IG}(a, b) \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Omega}) \quad (5a)$$

$$\text{vec}([\alpha, \mathbf{B}, \mathbf{A}_{1:L}]) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{V}_c) \quad \mathbf{Y}_t = \alpha + \mathbf{BZ}_t + \sum_{l=1}^L \mathbf{A}_l \mathbf{Y}_{t-l} + \varepsilon_t \quad (5b)$$

Ingredient 2: Neighbourhoods – Let \mathcal{S} be all locations and $N(s) \subseteq \mathcal{S}$ the neighbourhood of s . (E.g., stations $> 10\text{km}$ & $> 20\text{km}$ away)



Figure 1 – Neighbourhood graphs using road distances. **Left:** Complete graph. **Center:** stations $< 10\text{km}$ apart. **Right:** Stations $< 10\text{km}$ and $< 20\text{km}$ apart.



Extension I: Multivariate/Spatio-Temporal models

$$\mathbf{Y}_t = \boldsymbol{\alpha} + \mathbf{BZ}_t + \mathbf{A}_1 \mathbf{Y}_{t-1} + \mathbf{A}_2 \mathbf{Y}_{t-2} + \mathbf{A}_3 \mathbf{Y}_{t-3} + \varepsilon_t$$



Advantages:

- (1) **Sparse** and **interpretable** coefficient matrices \mathbf{A}_l
- (2) Theoretical guarantees for large class of stationary processes (Inoue et al., 2018, e.g.)
- (3) Allows fast on-line computation & **incremental updates**
 $[\mathcal{O}(\min\{p^3, S^3\})]$ for p regressors and spatial dimension S



Neighbourhoods set a priori \implies different nbhs via $\{m_1, \dots, m_K\}$



Extension II: Model Selection

Idea: We allow a change of models at CP locations

New Random Variable: m_t , the model at time t

$$r_t | r_{t-1} \sim H(r_t, r_{t-1}) \quad [\text{conditional CP prior}] \quad (6a)$$

$$m_t | m_{t-1}, r_t \sim q(m_t | m_{t-1}, r_t) \quad [\text{conditional model prior}] \quad (6b)$$

$$\theta_{m_t} | m_t \sim \pi_{m_t}(\theta_{m_t}) \quad [\text{parameter prior}] \quad (6c)$$

$$\mathbf{y}_t | m_t, \theta_{m_t} \sim f_{m_t}(\mathbf{y}_t | \theta_{m_t}) \quad [\text{observation density prior}] \quad (6d)$$

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

New Recursion:

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

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



Extension II: Model Selection

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

For the new term involving m_t , we have that

$$q(m_t | \mathbf{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) = \begin{cases} q(m_t) & \text{if } r_{t-1} = 0, \\ q(m_{t-1} | \mathbf{y}_{1:(t-1)}, r_{t-1}) & \text{if } r_{t-1} > 0. \end{cases} \quad (8)$$

with

$$q(m_{t-1} | \mathbf{y}_{1:(t-1)}, r_{t-1}) = \frac{p(\mathbf{y}_{1:(t-1)}, r_{t-1}, m_{t-1})}{\sum_{m_{t-1}} p(\mathbf{y}_{1:(t-1)}, r_{t-1}, m_{t-1})}. \quad (9)$$

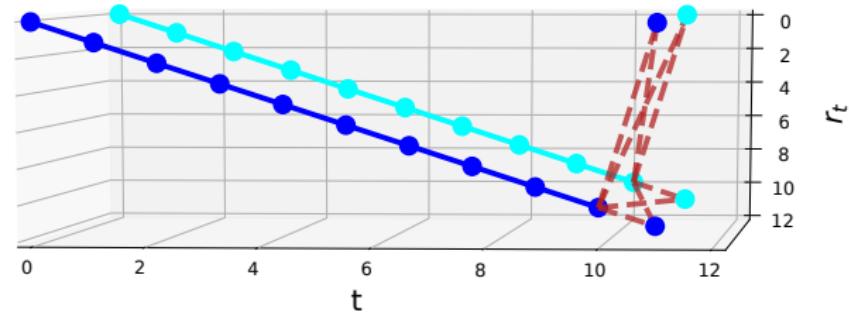
Interpretation: model posterior at time $t - 1$ becomes prior at t !



Extension II: Model Selection

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

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



Observations:

- (1) Each segment can now be described by a different model (f_m, π_m)
- (2) CPs are shifts in models m and/or their parameterizations θ_m
- (3) model prior at $t = 0$ is posterior at $t - 1 = q(m_t | \mathbf{y}_{1:(t-1)}, r_{t-1}, m_{t-1})$



Extension II: Model Selection

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

Inference:

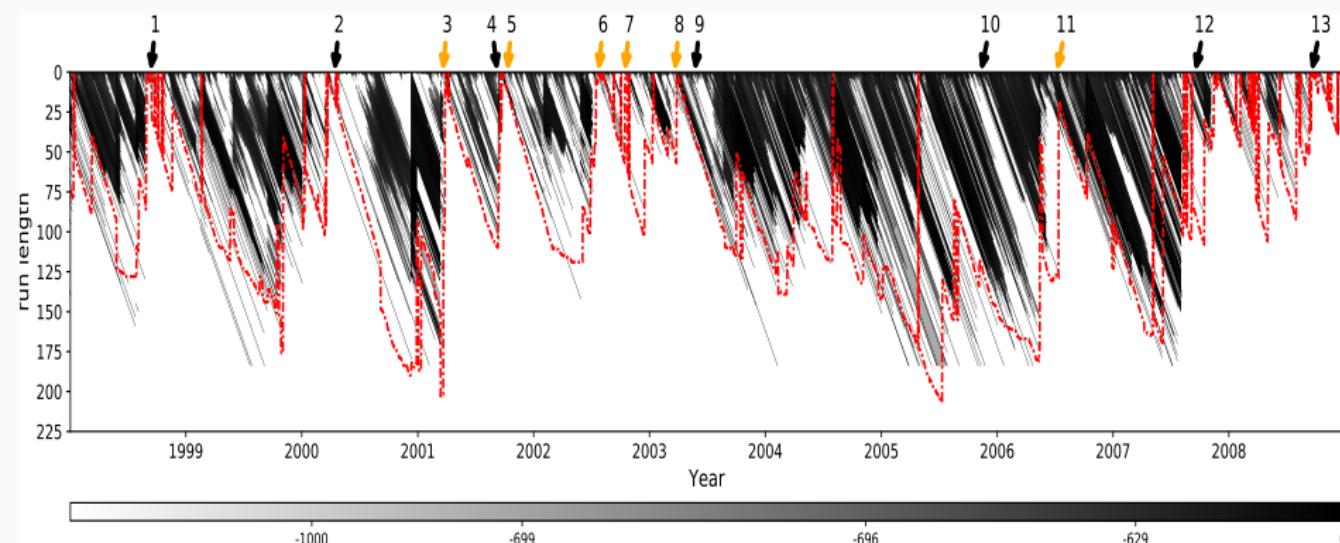
- (1) Evidence: $p(\mathbf{y}_{1:t}) = \sum_{r_t, m_t} p(\mathbf{y}_{1:t}, r_t, m_t)$
- (2) run-length & model posterior: $p(r_t, m_t | \mathbf{y}_{1:t}) = p(\mathbf{y}_{1:t}, r_t, m_t) / p(\mathbf{y}_{1:t})$
- (3) Prediction: $p(\mathbf{y}_{t+1} | \mathbf{y}_{1:t}) = \sum_{r_t, m_t} f_{m_t}(\mathbf{y}_{t+1} | \mathbf{y}_{1:t}, r_t) p(r_t, m_t | \mathbf{y}_{1:t})$
- (4) Run-length marginal posterior: $p(r_t | \mathbf{y}_{1:t}) = \sum_{m_t} p(r_t, m_t | \mathbf{y}_{1:t})$
- (5) Model marginal posterior: $p(m_t | \mathbf{y}_{1:t}) = \sum_{r_t} p(r_t, m_t | \mathbf{y}_{1:t}).$
- (6) MAP segmentation:

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



Improved CP detection [on Nile data]

✓ Improved CP detection [on 30 Portfolio return data]

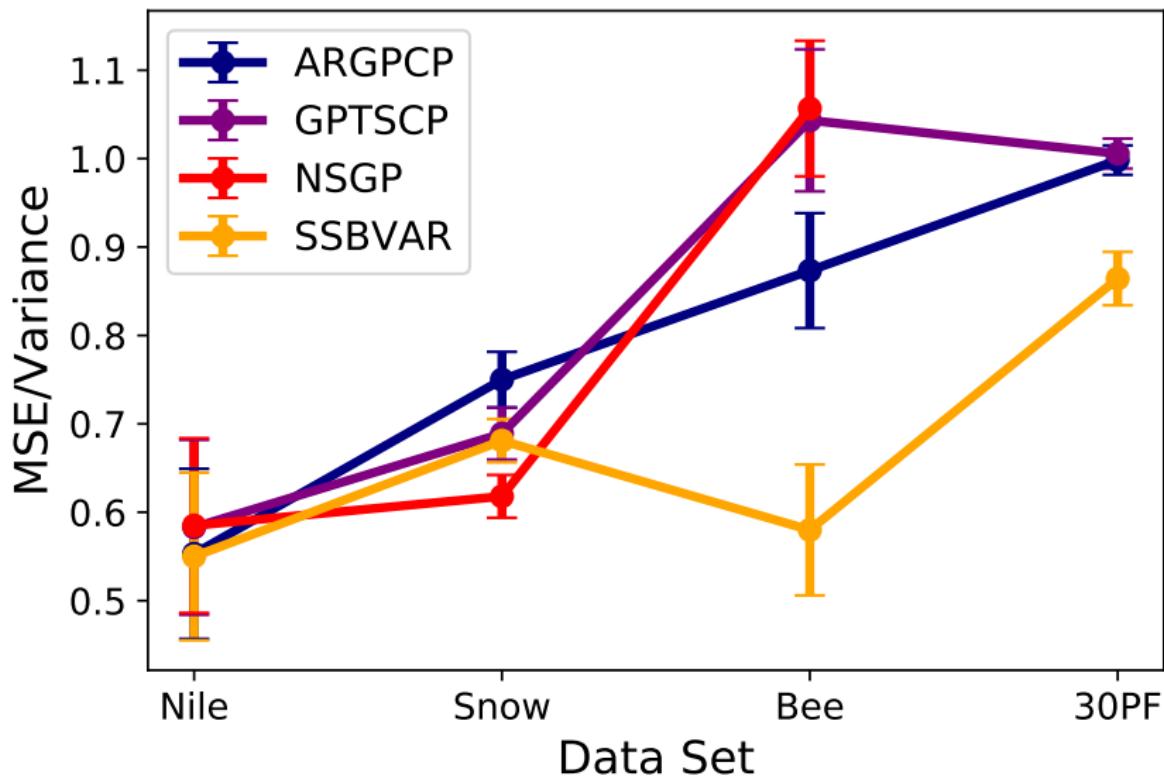


30 Portfolio return data, 01/01/1998–31/12/2008. CPs found with GPs by Saatci et al. (2010) in **black**, some **new CPs** found by BOCPDMS are:

- (3) OPEC cuts output by 4%,
- (8) Iraq war,
- (11) Iran announces successful enrichment of Uranium,

✓ Improved Prediction of multiple VARs vs single GP models

[MSE values for GP changepoint models as in Saatçi et al. (2010)]





Fast computation: E.g., much faster than GP CP-models

Theory:

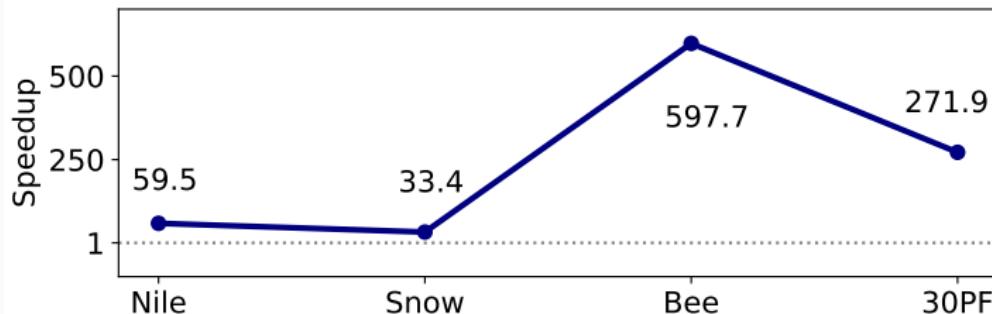
Processing $y_{1:T} \in \mathcal{R}^{T \times d}$ using a VAR, tracking K most likely (r_t, m_t) :

BOCPDMS: $\mathcal{O}\left(\sum_{t=1}^T K \cdot C_{p,d}\right) = \mathcal{O}(TKC_{p,d}) = \mathcal{O}(TK)$

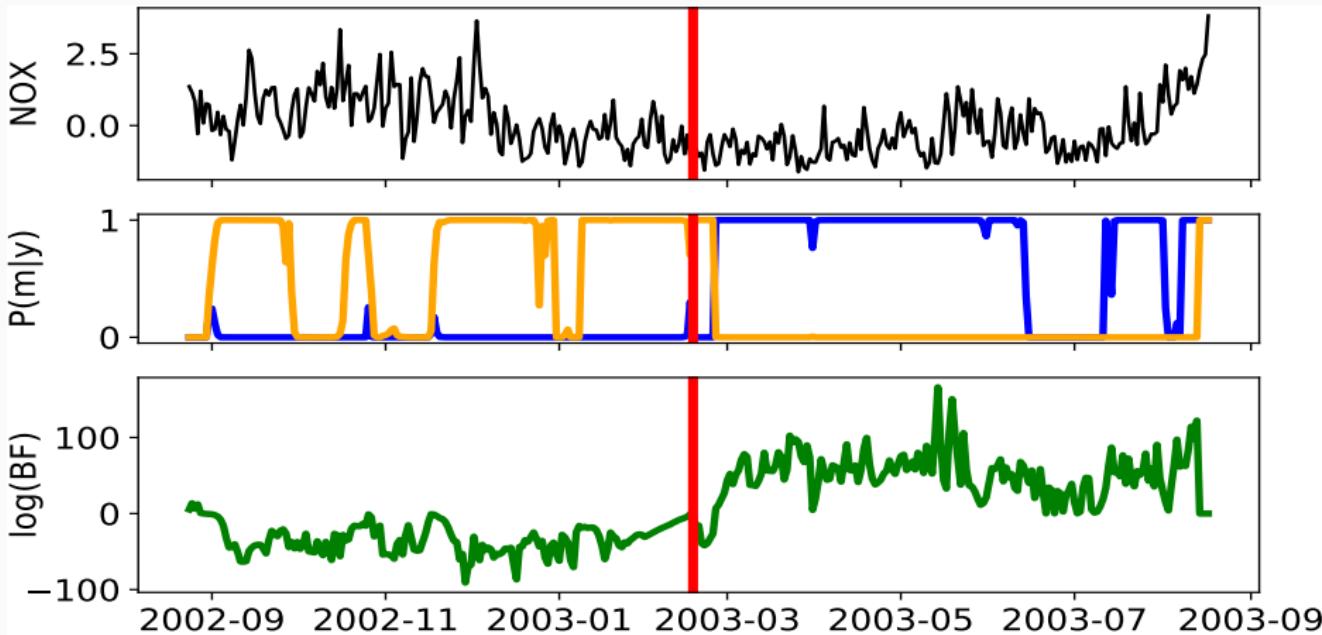
Previous GP CP models: $\mathcal{O}(TK^3)$

⇒ BOCPDMS using VARs is faster by a factor of K^2 .

Practice:



✓ New capability: 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

Summary: Novelty & Implications so far

Novel Features added to BOCPD:

-  Multivariate modelling of dependencies between data streams
-  Generalizing BOCPD to model selection

Practical Implications:

-  Improved CP detection
-  Improved Prediction, especially in multivariate data
-  New capability: Inference on shifting multivariate dynamics

Unresolved Limitation:

-  **non-robust** to outliers
-  **non-robust** to model misspecification



Limitation: BOCPD is not robust to outliers/misspecification

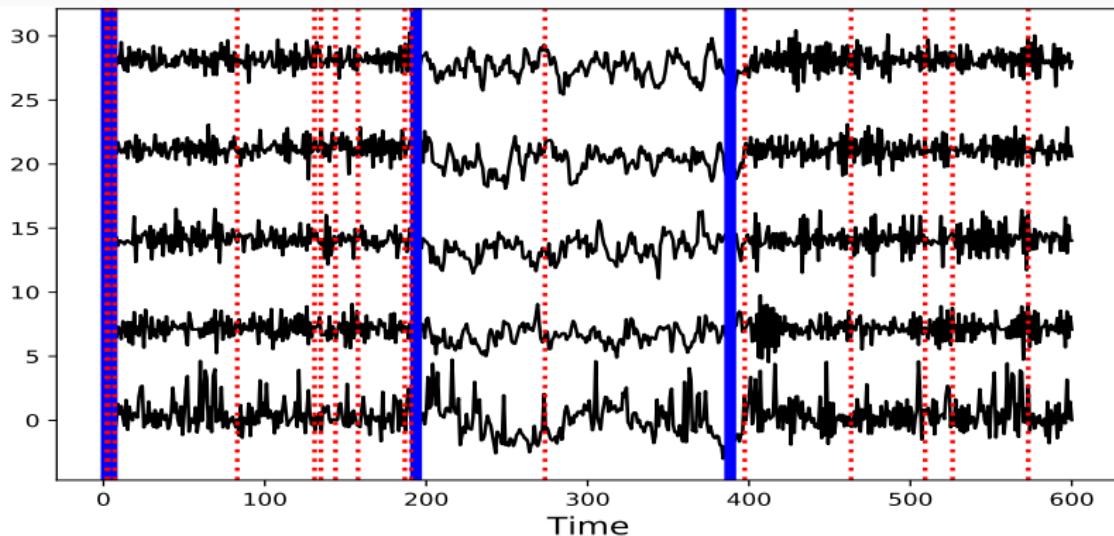


Figure 2 – E.g., 5-dimensional AR(1) with outliers in bottom-most series.

Segmentation we want and **segmentation we get** using BOCPD with AR(1).



Limitation: BOCPD is not robust to outliers/misspecification

How is it non-robust? (Superficial reasons)



On-line processing



Moderate/high dimensions for y_t

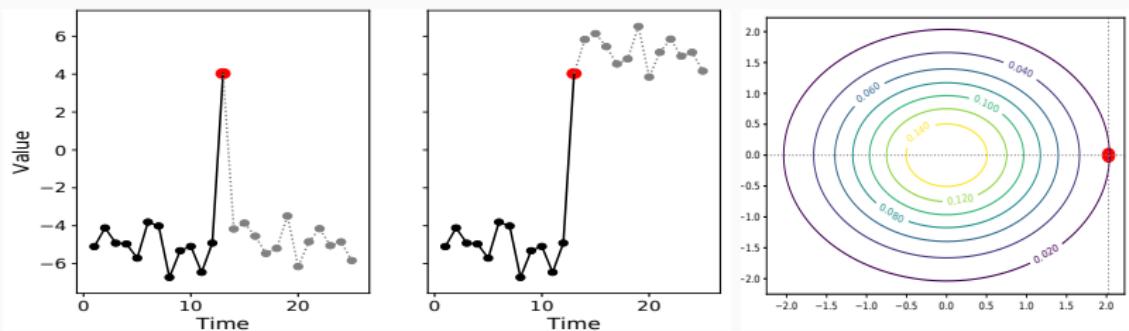


Figure 3 – Left, Center: Price for on-line processing is that outliers are confused with changepoints. **Right:** Multivariate densities become very small even if outliers occur only in a single dimension.



Limitation: BOCPD is not robust to outliers/misspecification

Why is it non-robust? (Fundamental reason for ⚡)

⚠ Influence function associated with (standard) Bayesian inference!

🔍 Generalized Bayesian Inference with robust divergences!

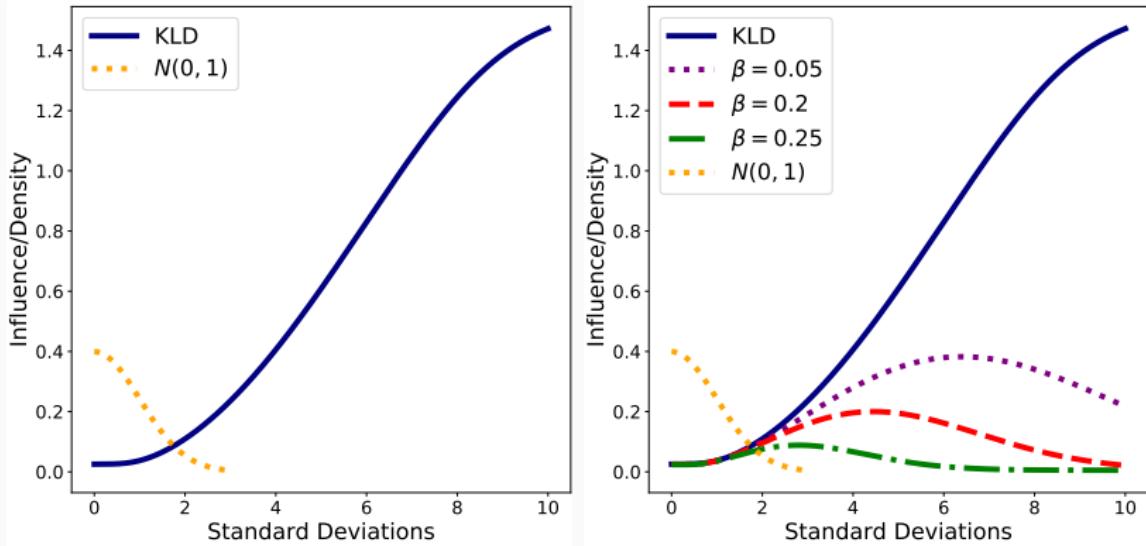


Figure 4 – Left: standard Bayes influence (Kullback-Leibler Divergence (KLD)) and standard normal density. **Right:** Robust β -Divergence (β -D) family.



Extension III: Robustification

We propose Generalized Bayesian Inference (GBI) Bissiri et al. (2016) to make BOCPD **doubly robust** for inference on θ_m and on (r_t, m_t) .

Idea of GBI: Bayesian belief updates about the parameter θ_m driven by arbitrary loss functions $\ell(\theta_m | \mathbf{y}_i)$:

$$\pi_m^{\text{GBI}}(\theta_m | \mathbf{y}_{(t-r_t):t}) \propto \pi_m(\theta) \exp \left\{ -\sum_{i=t-r_t}^t \ell(\theta_m | \mathbf{y}_i) \right\} \quad (11)$$

Problem: This is great if you want to do something like inference on the median (choose $\ell(\theta_m | \mathbf{y}_i) = \|\theta_m - \mathbf{y}_i\|_1$), but can you use this to do inference for dynamic probabilistic models ?

Solution: Work of Jewson et al. (2018) uses 1:1 correspondence between divergences and losses to propose generalized divergence-based GBI for a given probabilistic model. \implies Allows you to do robust Bayesian inference on arbitrary models



Extension III: Robustification

We propose β -D-based Generalized Bayesian Inference (GBI) to make BOCPD **doubly robust**: For the inference on θ_m and on (r_t, m_t) .

Parameter Layer robustified via β_p :

$$\pi_m^\beta(\theta_m | \mathbf{y}_{(t-r_t):t}) \propto \pi_m(\theta) \exp \left\{ -\sum_{i=t-r_t}^t \ell^\beta(\theta_m | \mathbf{y}_i) \right\}, \quad (12)$$

$$\ell^\beta(\theta_m | \mathbf{y}_t) = - \left(\frac{1}{\beta_p} f_m(\mathbf{y}_t | \theta_m)^{\beta_p} - \frac{1}{1 + \beta_p} \int_{\mathcal{Y}} f_m(\mathbf{z} | \theta_m)^{1 + \beta_p} d\mathbf{z} \right). \quad (13)$$

We then ensure that

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

is again available in closed form by forcing $\pi_m^\beta(\theta_m | \mathbf{y}_{(t-r_t):t})$ into conjugate form with $f_m(\mathbf{y}_t | \theta_m)$ via Structural Variational Inference.



Extension III: Robustification

We propose β -D-based Generalized Bayesian Inference (GBI) to make BOCPD **doubly robust**: For the inference on θ_m and on (r_t, m_t) .

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

$$\tilde{f}_m(\mathbf{y}_t | \mathbf{y}_{1:(t-1)}, r_{t-1}) = \underbrace{e^{-\left(\frac{1}{\beta_{rlm}} f_m(\mathbf{y}_t | \mathbf{y}_{1:(t-1)}, r_{t-1})^{\beta_{rlm}} - \frac{1}{1+\beta_{rlm}} \int_{\mathcal{Y}} f_m(\mathbf{z} | \mathbf{y}_{1:(t-1)}, r_{t-1})^{1+\beta_{rlm}} d\mathbf{z}\right)}}_{\text{=Same loss function } \ell^\beta, \text{ but this time for } (r_t, m_t)}$$

New Inference/Recursion:

$$p^\beta(\mathbf{y}_{1:t}, r_t, m_t) \propto \sum_{m_{t-1}, r_{t-1}} \left\{ \tilde{f}_{m_t}(\mathbf{y}_t | \mathbf{y}_{1:(t-1)}, r_{t-1}) q(m_t | \mathbf{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) H(r_t, r_{t-1}) p^\beta(\mathbf{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \right\}$$

Note: Nothing further needed to use this for inference exactly as before!



Extension III: Robustification

🔒 β -D posterior not scalable \Rightarrow 🔑 Structural Variational Inference

Observation I: As $\beta \rightarrow 0$, β -D \rightarrow KLD! $\Rightarrow \pi_m^{\text{KLD}} \approx \pi_m^\beta$ for small β !

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

$$\hat{\pi}_m^{\beta_p}(\theta_m) = \underset{\pi_m^{\text{KLD}}(\theta_m)}{\operatorname{argmin}} \left\{ \text{KL} \left(\pi_m^{\text{KLD}}(\theta_m) \middle\| \pi_m^{\beta_p}(\theta_m | \mathbf{y}_{(t-r_t):t}) \right) \right\}. \quad (14)$$

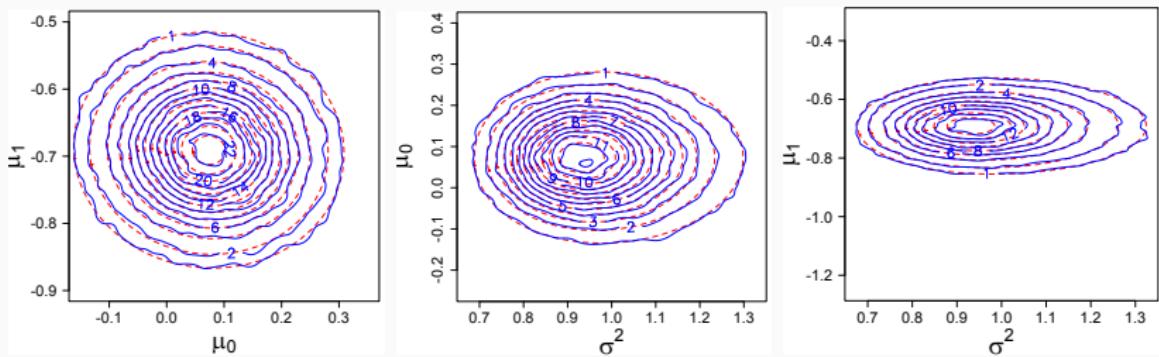


Figure 5 – Contour plots of bivariate marginals of approximation $\hat{\pi}_m^{\beta_p}(\theta_m)$ (dashed) and the target $\pi_m^{\beta_p}(\theta_m | \mathbf{y}_{(t-r_t):t})$ (solid) estimated from 95,000 Hamiltonian Monte Carlo samples for BLR ($d = 1$, two regressors, $\beta_p = 0.25$).



Extension III: Robustification

🔒 Choice of $\beta \implies$ 🔑 Initialization.

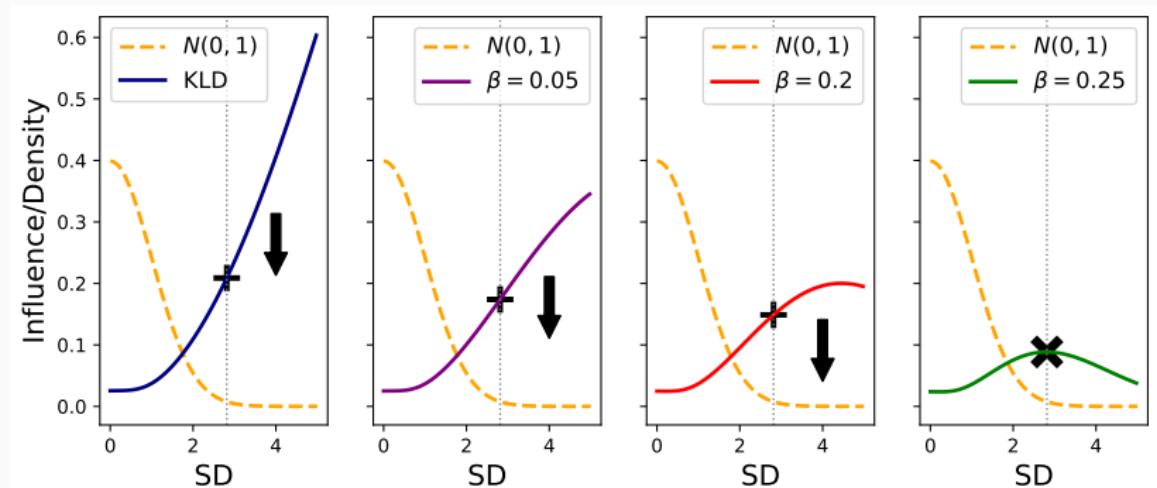


Figure 6 – Initialization procedure visualized for a (standard) normal prior on $y_t \in \mathbb{R}$ and if we want to have maximum influence of an observation 2.75 standard deviation units from the expected value under the current belief.



Extension III: Robustification

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

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

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

One can then minimize L on-line via

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - \eta \cdot \begin{bmatrix} \nabla_{\beta_{\text{rlm}, t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right) \\ \nabla_{\beta_{p, t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right) \end{bmatrix} \quad (15)$$

New capability: CP detection in outlier-prone data streams

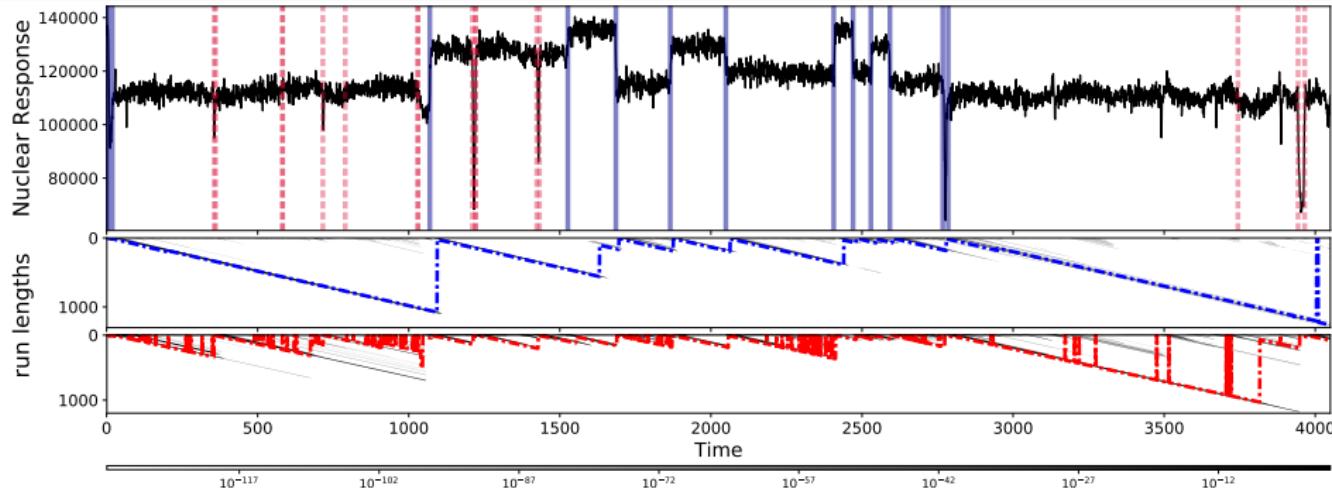


Figure 7 – Robust segmentation and run-length distribution and additionally found CPs with non-robust run-length distribution

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

Better Model Selection on shifting multivariate dynamics

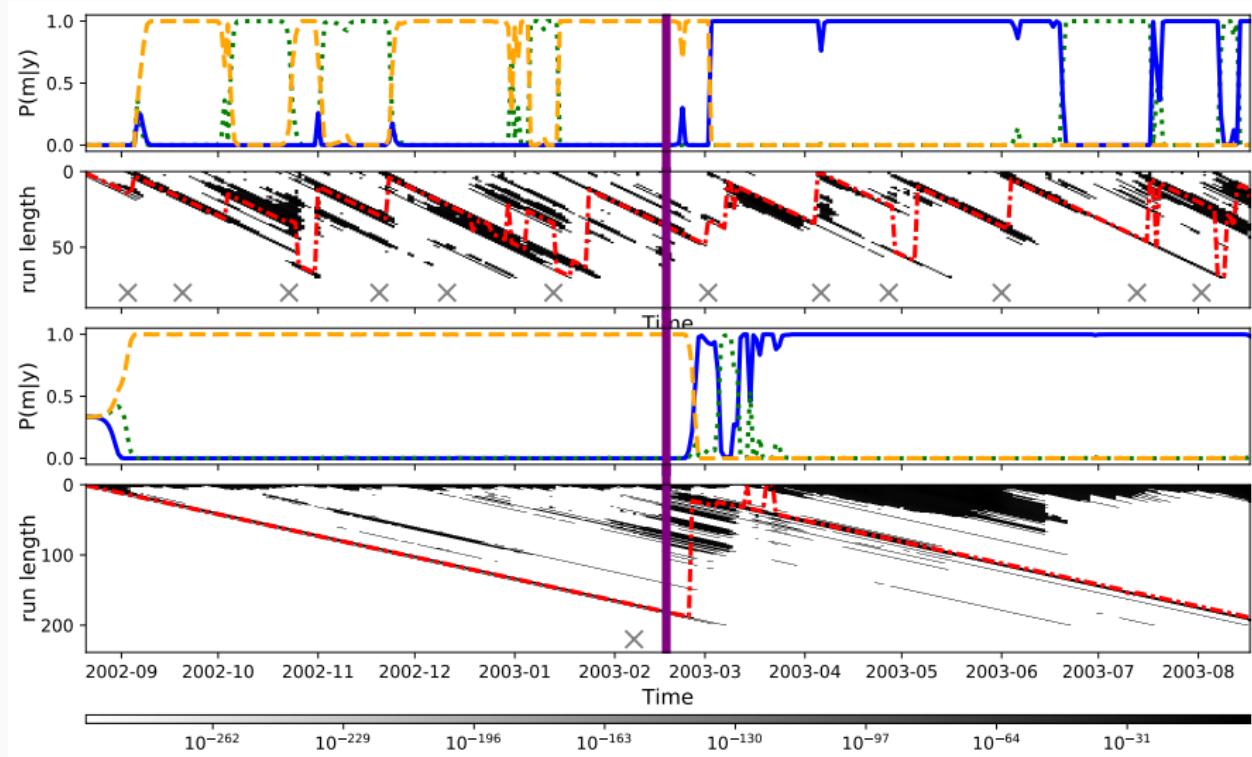


Figure 8 – Top & bottom two panels: standard & robust BOCPD.

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.
- Inoue, A., Kasahara, Y., Pourahmadi, M., et al. (2018). Baxter’s inequality for finite predictor coefficients of multivariate long-memory stationary processes. *Bernoulli*, 24(2):1202–1232.
- 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)*. to appear.
- 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)*. submitted to.
- Saatçi, Y., Turner, R. D., and Rasmussen, C. E. (2010). Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934.