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

Jeremias Knoblauch
The Alan Turing Institute
Department of Statistics
University of Warwick
Coventry, CV4 7AL
j.knoblauch@warwick.ac.uk

Jack Jewson
Department of Statistics
University of Warwick
Coventry, CV4 7AL
j.e.jewson@warwick.ac.uk

Theodoros Damoulas
The Alan Turing Institute
Department of Computer Science & Department of Statistics
University of Warwick
Coventry, CV4 7AL
t.damoulas@warwick.ac.uk

Abstract

We present the very first robust Bayesian Online Changepoint Detection algorithm through General Bayesian Inference (GBI) with β-divergences. The resulting inference procedure is doubly robust for both the parameter and the changepoint (CP) posterior, with linear time and constant space complexity. We provide a construction for exponential models and demonstrate it on the Bayesian Linear Regression model. In so doing, we make two additional contributions: Firstly, we make GBI scalable using Structural Variational approximations that are exact as $\beta \to 0$. Secondly, we give a principled way of choosing the divergence parameter β by minimizing expected predictive loss on-line. Reducing False Discovery Rates of CPs from up to 99% to 0% on real world data, this offers the state of the art.

1 Introduction

Modeling non-stationary time series with changepoints (CPs) is popular [23, 50, 33] and important in a wide variety of research fields, including genetics [8, 16, 42], finance [27], oceanography [24], brain imaging and cognition [13, 20], cybersecurity [37] and robotics [2, 26]. For streaming data, a particularly important subclass are Bayesian On-line Changepoint Detection (BOCPD) methods that can process data sequentially [1, 11, 43, 47, 46, 41, 8, 34, 44, 40, 25] while providing full probabilistic uncertainty quantification. These algorithms declare CPs if the posterior predictive computed from $\boldsymbol{y}_{1:t}$ at time t has low density for the value of the observation $\boldsymbol{y}_{t+1}$ at time $t+1$. Naturally, this leads to a high false CP discovery rate in the presence of outliers and as they run on-line, pre-processing is not an option. In this work, we provide the first robust on-line CP detection method that is applicable to multivariate data, works with a class of scalable models and quantifies model, CP and parameter uncertainty in a principled Bayesian fashion.

Standard Bayesian inference minimizes the Kullback-Leibler divergence (KLD) between the fitted model and the Data Generating Mechanism (DGM), but is not robust under outliers or model misspecification due to its strictly increasing influence function. We remedy this by instead minimizing the β-divergence (β-D) whose influence function has a unique maximum, allowing us to deal with outliers effectively. Fig. 1 **A** illustrates this: Under the β-D, the influence of observations first

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

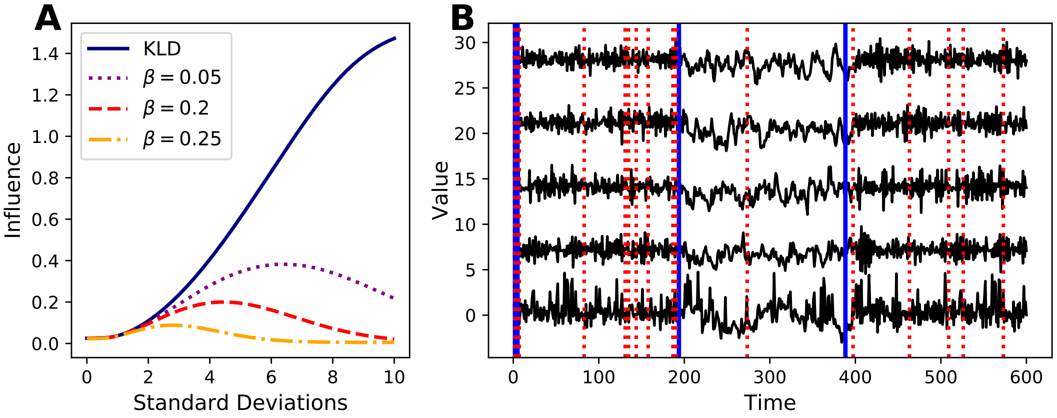

Figure 1: **A:** Influence of $\boldsymbol{y}_t$ on inference as function of distance to the posterior expectation in Standard Deviations for β-divergences with different βs. **B:** Five jointly modeled Simulated Autoregressions (ARs) with true CPs at $t = 200, 400$; bottom-most AR injected with t_4-noise. Maximum A Posteriori CPs of robust (standard) BOCPD shown as solid (dashed) vertical lines.

increases as they move away from the posterior mean, mimicking the KLD. However, once they move far enough, their influence decreases again. This can be interpreted to mean that they are (increasingly) treated as outliers. As β increases, observations are registered as outliers closer to the posterior mean. Conversely, as $\beta \to 0$, one recovers the KLD which cannot treat any observation as an outlier. In addressing misspecification and outliers this way, our approach builds on the principles of General Bayesian Inference (GBI) [see 6, 21] and robust divergences [e.g. 4, 15]. This paper presents three contributions in separate domains that are also illustrated in Figs. 1 and 3:

(1) **Robust BOCPD**: We construct the very first robust BOCPD inference. The procedure is applicable to a wide class of (multivariate) models and is demonstrated on Bayesian Linear Regression (BLR). Unlike standard BOCPD, it discerns outliers and CPs, see Fig. 1 **B**.

(2) **Scalable GBI:** Due to intractable posteriors, GBI has received little attention in machine learning so far. We remedy this with a Structural Variational approximation which preserves parameter dependence and is exact as $\beta \to 0$, providing a near-perfect fit, see Fig. 3.

(3) **Choosing β:** While Fig. 1 **A** shows that β regulates the degree of robustness [see also 21, 15], it is unclear how to set its magnitude. For the first time, we provide a principled way of initializing β. Further, we show how to refine it on-line by minimizing predictive losses.

The remainder of the paper is structured as follows: In Section 2, we summarize standard BOCPD and show how to extend it to robust inference using the β-D. We quantify the degree of robustness and show that inference under the β-D can be designed so that a single outlier never results in false declaration of a CP, which is impossible under the KLD. Section 3 motivates efficient Structural Variational Inference (SVI) with the β-D posterior. Within BOCPD, we propose to scale SVI using variance-reduced Stochastic Gradient Descent. Next, Section 4 expands on how β can be initialized before the algorithm is run and then optimized on-line during execution time. Lastly, Section 5 showcases the substantial gains in performance of robust BOCPD when compared to its standard version on real world data in terms of both predictive error and CP detection.

2 Using Bayesian On-line Changepoint Detection with β-Divergences

BOCPD is based on the Product Partition Model [3] and introduced independently in Adams and MacKay [1] and Fearnhead and Liu [11]. Recently, both formulations have been unified in Knoblauch and Damoulas [25]. The underlying algorithm has extensions ranging from Gaussian Processes [41] and on-line hyperparameter optimization [8] to non-exponential families [44, 34].

To formulate BOCPD probabilistically, define the run-length r_t as the number of observations at time t since the most recent CP and m_t as the best model in the set $\mathcal{M}$ for the observations since that CP. Then, given a real-valued multivariate process $\{\boldsymbol{y}_t\}_{t=1}^{\infty}$ of dimension d, a model universe $\mathcal{M}$, a run-length prior h defined over $\mathbb{N}_0$ and a model prior q over $\mathcal{M}$, the BOCPD model is

$$r_t|r_{t-1} \sim H(r_t, r_{t-1}) \qquad m_t|m_{t-1}, r_t \sim q(m_t|m_{t-1}, r_t) \tag{1a}$$

$$\boldsymbol{\theta}_m|m_t \sim \pi_{m_t}(\boldsymbol{\theta}_{m_t}) \qquad \boldsymbol{y}_t|m_t, \boldsymbol{\theta}_{m_t} \sim f_{m_t}(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t}) \tag{1b}$$

where $q(m_t|m_{t-1}, r_t) = 1_{m_{t-1}}(m_t)$ for $r_t > 0$ and $q(m_t)$ otherwise, and where H is the conditional run-length prior so that $H(0, r) = h(r+1)$, $H(r+1, r) = 1-h(r+1)$ for any $r \in \mathbb{N}_0$ and $H(r, r') = 0$ otherwise. For example, Bayesian Linear Regression (BLR) with the $d \times p$ regressor matrix $\boldsymbol{X}_t$ and prior covariance Σ_0 is given by $\boldsymbol{\theta}_m = (\sigma^2, \boldsymbol{\mu})$, $f_m(\boldsymbol{y}_t|\boldsymbol{\theta}_m) = \mathcal{N}_d(\boldsymbol{y}_t; \boldsymbol{X}_t\boldsymbol{\mu}, I_d)$ and $\pi_m(\boldsymbol{\theta}_m) = \mathcal{N}_d(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \sigma^2\Sigma_0)\mathcal{IG}(\sigma^2; a_0, b_0)$. If the computations of the parameter posterior $\pi_m(\boldsymbol{\theta}_m|\boldsymbol{y}_{1:t}, r_t)$ and the posterior predictive $f_m(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_t) = \int_{\Theta_m} f_m(\boldsymbol{y}_t|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m|\boldsymbol{y}_{1:(t-1)}, r_t)d\boldsymbol{\theta}_m$ are efficient for all models $m \in \mathcal{M}$, then so is the recursive computation given by

$$p(\boldsymbol{y}_1, r_1 = 0, m_1) = q(m_1) \cdot \int_{\Theta_{m_1}} f_{m_1}(\boldsymbol{y}_1|\boldsymbol{\theta}_{m_1})\pi_{m_1}(\boldsymbol{\theta}_{m_1})d\boldsymbol{\theta}_{m_1} = q(m_1) \cdot f_{m_1}(\boldsymbol{y}_1|\boldsymbol{y}_0), \tag{2a}$$

$$p(\boldsymbol{y}_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \Big\{ f_{m_t}(\boldsymbol{y}_t|\mathcal{F}_{t-1})q(m_t|\mathcal{F}_{t-1}, m_{t-1})H(r_t, r_{t-1})p(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\} \tag{2b}$$

where $\mathcal{F}_{t-1} = \{\boldsymbol{y}_{1:(t-1)}, r_{t-1}\}$ and $p(\boldsymbol{y}_{1:t}, r_t, m_t)$ is the joint density of $\boldsymbol{y}_{1:t}$, m_t and r_t. The run-length and model posteriors are then available exactly at time t, as $p(r_t, m_t|\boldsymbol{y}_{1:t}) = p(\boldsymbol{y}_{1:t}, r_t, m_t)/\sum_{m_t, r_t} p(\boldsymbol{y}_{1:t}, r_t, m_t)$. For a full derivation and the resulting inference see [25].

2.1 General Bayesian Inference (GBI) with β-Divergences (β-D)

Standard Bayesian inference minimizes the KLD between the Data Generating Mechanism (DGM) and its probabilistic model (see Section 2.1 of [6] for a clear illustration). In the M-closed world where one assumes that the DGM and model coincide, the KLD is the most efficient way of updating posterior beliefs. However, this is no longer the case in the M-open world [5] where they match only approximately [21], e.g. in the presence of outliers. GBI [6, 21] generalizes standard Bayesian updating based on the KLD to a family of divergences. In particular, it uses the relationship between losses ℓ and divergences D to deduce for D a corresponding loss ℓ^D. It can then be shown that for model m, the posterior update optimal for D yields the distribution

$$\pi_m^D(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \propto \pi_m(\theta) \exp\left\{-\sum_{i=t-r_t}^{t} \ell^D(\boldsymbol{\theta}_m|\boldsymbol{y}_i)\right\}. \tag{3}$$

For parameter inference with the KLD and β-D, these losses are the log score and the Tsallis score:

$$\ell^{\text{KLD}}(\boldsymbol{\theta}_m|\boldsymbol{y}_t) = -\log\left(f_m(\boldsymbol{y}_t|\boldsymbol{\theta}_m)\right. \tag{4}$$

$$\ell^{\beta}(\boldsymbol{\theta}_m|\boldsymbol{y}_t) = -\left(\frac{1}{\beta_{\text{p}}} f_m(\boldsymbol{y}_t|\boldsymbol{\theta}_m)^{\beta_{\text{p}}} - \frac{1}{1+\beta_{\text{p}}}\int_{\mathcal{Y}} f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)^{1+\beta_{\text{p}}} d\boldsymbol{z}\right). \tag{5}$$

Eq. (5) shows why the β-D excels at robust inference: Similar to tempering, ℓ^β exponentially downweights the density, attaching less influence to observations in the tails of the model. This phenomenon is depicted with influence functions $I(\boldsymbol{y}_t)$ in Figure 1 **A**. $I(\boldsymbol{y}_t)$ is a divergence between the posterior with and without an observation $\boldsymbol{y}_t$ [28].

GBI with the β-D yields robust inference without the need to specify a heavy-tailed or otherwise robustified model. Hence, one estimates the same model parameters as in standard Bayesian inference while down-weighting the influence of observations that are overly inconsistent with the model. Accordingly, GBI provides robust inference for a much wider class of models and situations than the ones illustrated here. Though other divergences such as α-Divergences [e.g. 19] also accommodate robust inference, we restrict ourselves to the β-D. We do this because unlike other divergences, it does not require estimation of the DGM's density. Density estimation increases estimation error, is computationally cumbersome and works poorly for small run-lengths (i.e. sample sizes). Note that versions of GBI have been proposed before [14, 32, 38, 10], but have framed the procedure as alternative to Variational Bayes instead.

Apart from the computational gains of Section 3.1, we tackle robust inference via the β-D rather than via Student's t errors for three reasons: Firstly, robust run-length posteriors need robustness in *ratios* rather than *tails* (see Section 2.3 and the simulation results for Student's t errors in the Appendix). Secondly, Student's t errors model outliers as part of the DGM, which compromises

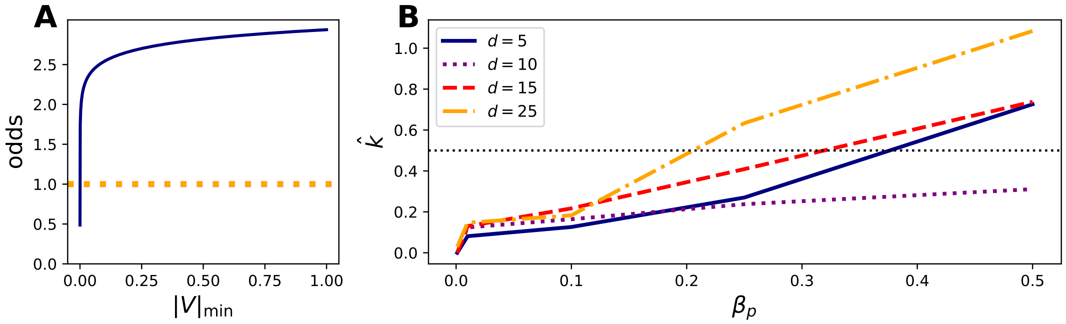

Figure 2: **A**: Lower bound on the odds of Thm. 1 for priors used for Figure 1 **B** and $h(r) = 1/100$. **B**: $\hat{k}$ for different choices of β_{p} and output (input) dimensions d ($2d$) in an autoregressive BLR

the inference target: Consider a BLR with error $e_t = \varepsilon_t + w_t\nu_t$, where $w_t \sim \text{Ber}(p)$ for $p = 0.01$, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with outliers $\nu_t \sim t_1(0, \gamma)$. Appropriate choices of β_{p} give most influence to the $(1-p)\cdot 100\% = 99\%$ of typical observations one can explain well with the BLR model. In contrast, modeling e_t as Student's t under the KLD lets ν_t dominate parameter inference and lets 1% of observations inflate the predictive variance substantially. Thirdly, using Student's t errors is a technique only applicable to symmetric, continuous models. In contrast, GBI with the β-D is valid for any setting, e.g. for asymmetric errors as well as point and count processes.

2.2 Robust BOCPD

The literature on robust on-line CP detection so far is sparse and covers limited settings without Bayesian uncertainty quantification [e.g. 36, 7, 12]. For example, the method in Fearnhead and Rigaill [12] only produces point estimates and is limited to fitting a piecewise constant function to univariate data. In contrast, BOCPD can be applied to multivariate data and a set of models $\mathcal{M}$ while quantifying uncertainty about these models, their parameters and potential CPs, but is not robust. Noting that for standard BOCPD the posterior expectation is given by

$$\mathbb{E}\left(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}\right) = \sum_{r_t, m_t} \mathbb{E}\left(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}\right) p(r_{t-1}, m_{t-1}|\boldsymbol{y}_{1:(t-1)}), \tag{6}$$

the key observation is that prediction is driven by two probability distributions: The run-length and model posterior $p(r_t, m_t|\boldsymbol{y}_{1:t})$ and parameter posterior distributions $\pi_m(\boldsymbol{\theta}_m|\boldsymbol{y}_{1:t})$. Thus, we make BOCPD robust by using β-D posteriors $p^{\beta_{\text{rlm}}}(r_t, m_t|\boldsymbol{y}_{1:t})$, $\pi_m^{\beta_{\text{p}}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{1:t})$ for $\boldsymbol{\beta} = (\beta_{\text{rlm}}, \beta_{\text{p}}) > 0$[1].

β_{rlm} prevents abrupt changes in $p^{\beta_{\text{rlm}}}(r_t, m_t|\boldsymbol{y}_{1:t})$ caused by a small number of observations, see section 2.3. This form of robustness is easy to implement and retains the closed forms of BOCPD: In Eqs. (2a) and (2b), one simply replaces $f_{m_t}(\boldsymbol{y}_t|\boldsymbol{y}_0)$ and $f_{m_t}(\boldsymbol{y}_t|\mathcal{F}_{t-1})$ by their β-D-counterparts $\exp\{\ell^{\beta_{\text{rlm}}}(\boldsymbol{\theta}_{m_t}|\boldsymbol{y}_t)\}$, where

$$\ell^{\beta_{\text{rlm}}}(\boldsymbol{\theta}_{m_t}|\boldsymbol{y}_t) = -\left(\frac{1}{\beta_{\text{rlm}}} f_m(\boldsymbol{y}_t|\mathcal{F}_{t-1})^{\beta_{\text{rlm}}} - \frac{1}{1+\beta_{\text{rlm}}}\int_{\mathcal{Y}} f_m(\boldsymbol{z}|\mathcal{F}_{t-1})^{1+\beta_{\text{rlm}}} d\boldsymbol{z}\right). \tag{7}$$

While the posterior $p^{\beta_{\text{rlm}}}(r_t, m_t|\boldsymbol{y}_{1:t})$ is only available up to a constant, it is discrete and thus easy to normalize. Complementing this, β_{p} regulates the robustness of $\pi_m^{\beta_{\text{p}}}(\boldsymbol{\theta}|\boldsymbol{y}_{1:t})$ by preventing it from being dominated by tail events. Section 3.1 overcomes the intractability of $\pi_m^{\beta_{\text{p}}}(\boldsymbol{\theta}|\boldsymbol{y}_{1:t})$ using Structural Variational Inference (SVI) that recovers the approximated distribution exactly as $\beta_{\text{p}} \to 0$.

2.3 Quantifying robustness

The algorithm of Fearnhead and Rigaill [12] is robust because hyperparameters enforce that a single outlier is insufficient for declaring a CP. Analogously, we investigate conditions under which a single (outlying) observation $\boldsymbol{y}_{t+1}$ is able to force a CP. An intuitive way of achieving this is by studying the odds of $r_{t+1} \in \{0, r+1\}$ conditional on $r_t = r$:

$$\frac{p(r_{t+1} = r+1|\boldsymbol{y}_{1:t+1}, r_t = r, m_t)}{p(r_{t+1} = 0|\boldsymbol{y}_{1:t+1}, r_t = r, m_t)} = \frac{p(\cancel{\boldsymbol{y}_{1:t}, r_t = r, m_t}) \cdot (1 - H(r_{t+1}, r_t)) f_{m_t}^D(\boldsymbol{y}_{t+1}|\mathcal{F}_t)}{p(\cancel{\boldsymbol{y}_{1:t}, r_t = r, m_t}) \cdot H(r_{t+1}, r_t) f_{m_t}^D(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0)}. \tag{8}$$

[1] In fact, β_{p}= β_p^m, i.e. the robustness is model-specific, but this is suppressed for readability

Here, $f^D_{m_t}$ denotes the negative exponential of the score under divergence D. In particular, $f^{\text{KLD}}_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t) = f_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$ and $f^{\beta_{\text{rlm}}}_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t) = \exp\left\{-\ell^{\beta_{\text{rlm}}}(\boldsymbol{\theta}_m|\boldsymbol{y}_t)\right\}$ as in Eq. (7). Taking a closer look at Eq. (8), if $\boldsymbol{y}_{t+1}$ is an outlier with low density under $f^D_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$, the odds will move in favor of a CP provided that the prior is sufficiently uninformative to make $f^D_{m_t}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0) > f^D_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$. In fact, even very small differences have a substantial impact on the odds. This is why using the Student's t error for the BLR model with standard Bayes will not provide robust run-length posteriors: While an outlying observation $\boldsymbol{y}_{t+1}$ will have greater density $f^{\text{KLD}}_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$ under a Student's t error model than under a normal error model, $f^{\text{KLD}}_{m_t}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0)$ (the density under the prior) will also be larger under the Student's t error model. As a result, changing the tails of the model only has a very limited effect on the ratio in Eq. (8). In fact, the perhaps unintuitive consequence is that Student's t error models will yield CP inference that very closely resembles that of the corresponding normal model. A range of numerical examples in the Appendix illustrate this surprising fact. In contrast, CP inference robustified via the β-D does not suffer from this phenomenon. In fact, Theorem 1 provides very mild conditions for the β-D robustified BLR model ensuring that the odds never favor a CP after *any* single outlying observation $\boldsymbol{y}_{t+1}$.

Theorem 1. If m_t in Eq. (8) is the Bayesian Linear Regression (BLR) model with $\boldsymbol{\mu} \in \mathbb{R}^p$ and priors a_0, b_0, μ_0, Σ_0; and if the posterior predictive's variance determinant is larger than $|V|_{\min} > 0$, then one can choose any $(\beta_{\text{rlm}}, H(r_t, r_{t+1})) \in S\left(p, \beta_{\text{rlm}}, a_0, b_0, \mu_0, \Sigma_0, |V|_{\min}\right)$ to guarantee that

$$\frac{(1 - H(r_{t+1}, r_t))f^{\beta_{\text{rlm}}}_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)}{H(r_{t+1}, r_t)f^{\beta_{\text{rlm}}}_{m_t}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0)} \geq 1, \tag{9}$$

where the set $S\left(p, \beta_{\text{rlm}}, a_0, b_0, \mu_0, \Sigma_0, |V|_{\min}\right)$ is defined by an inequality given in the Appendix.

Thm. 1 says that one can bound the odds for a CP independently of $\boldsymbol{y}_{t+1}$. The requirement for a lower bound $|V|_{\min}$ results from the integral term in Eq. (5), which dominates β-D-inference if $|V|$ is extremely small. In practice, this is not restrictive: E.g. for $p = 5$, $h(r) = \frac{1}{\lambda}$, $a_0 = 3, b_0 = 5, \Sigma_0 = \text{diag}(100, 5)$ used in Fig. 1 **B**, Thm. 1 holds for $(\beta_{\text{rlm}}, \lambda) = (0.15, 100)$ used for inference if $|V|_{\min} \geq 8.12 \times 10^{-6}$. Fig. 2 **A** plots the lower bound (see Appendix) as function of $|V|_{\min}$.

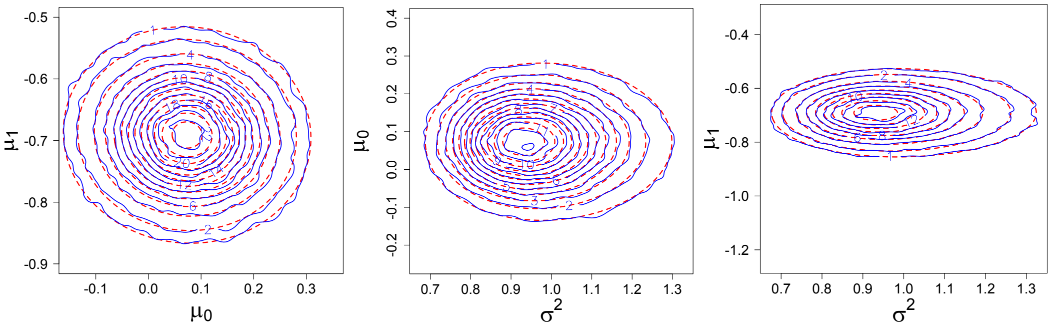

Figure 3: Exemplary contour plots of bivariate marginals for the approximation $\widehat{\pi}^{\beta_{\text{p}}}_m(\boldsymbol{\theta}_m)$ of Eq. (11) (dashed) and the target $\pi^{\beta_{\text{p}}}_m(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t})$ (solid) estimated and smoothed from 95,000 Hamiltonian Monte Carlo samples for the β-D posterior of BLR with $d = 1$, two regressors and $\beta_{\text{p}} = 0.25$.

3 On-line General Bayesian Inference (GBI)

3.1 Structural Variational Approximations for Conjugate Exponential Families

While there has been a recent surge in theoretical work on GBI [6, 15, 21, 14], applications have been sparse, in large part due to intractability. While sampling methods have been used successfully for GBI [21, 15], it is not easy to scale these for the robust BOCPD setting. Thus, most work on BOCPD has focused on conjugate distributions [1, 43, 11] and approximations [44, 34]. We extend the latter branch of research by deploying Structural Variational Inference (SVI). Unlike mean-field approximations, this preserves parameter dependence in the posterior, see Figure 3. While it is in principle possible to solve the inference task by sampling, this is computationally burdensome and makes the algorithm on-line in name only: Any sampling approach needs to (I) sample from $\pi^{\beta_p}_m(\theta_m|\boldsymbol{y}_{t-r_t:t})$ in Eq. (3), (II) numerically integrate to obtain $f_m(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_t)$ and lastly (III) sample and numerically integrate the integral in Eq. (7) which no longer has a closed form. Moreover, this has to be performed for each (r_t, m) at times $t = 1, 2, \ldots$. On top of this increased computational cost, it creates three sources of approximation error propagated forward through time via Eqs. (2a) and (2b). Since π_m^{KLD} is available in closed form and as β-D $\rightarrow$ KLD as $\beta \rightarrow 0$ [4], there is an especially compelling way of doing SVI for conjugate models using the β-D based on the fact that

$$\pi_m^{\beta_\text{p}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \approx \pi_m^{\text{KLD}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \tag{10}$$

is exact as $\beta \rightarrow 0$. Thus we approximate the β-D posterior for model m and run-length r_t as

$$\widehat{\pi}_m^{\beta_\text{p}}(\boldsymbol{\theta}_m) = \underset{\pi_m^{\text{KLD}}(\boldsymbol{\theta}_m)}{\operatorname{argmin}} \left\{ \text{KL}\left(\pi_m^{\text{KLD}}(\boldsymbol{\theta}_m) \middle\| \pi_m^{\beta_\text{p}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \right) \right\}. \tag{11}$$

While this ensures that the densities $\widehat{\pi}_m^{\beta_\text{p}}$ and π_m^{KLD} belong to the same family, the variational parameters can be very different from those implied by the KLD-posterior. This approximation mitigates multiple issues that would arise with sampling approaches: By forcing $\pi_m^{\beta_p}(\theta_m|\boldsymbol{y}_{1:t})$ into the conjugate closed form, steps (II) and (III) are solved analytically. Thus, inference is orders of magnitude faster, while the resulting approximation error remains negligible (see Figs 2B, 3).

Moreover, for many models, the Evidence Lower Bound (ELBO) associated with the optimization in Eq. (11) is available in closed form. As a result, off-the-shelf optimizers are sufficient and no black-box or sampling-based techniques are required to efficiently tackle the problem. Theorem 2 provides the conditions for a conjugate exponential family to admit such a closed form ELBO. The proof alongside the derivation of the ELBO for BLR can be found in the Appendix

Theorem 2. The ELBO objective corresponding to the β-D posterior approximation in Eq. (11) of an exponential family likelihood model $f_m(\boldsymbol{y}; \theta_m) = \exp\left(\eta(\theta_m)^T T(\boldsymbol{y})\right) g(\eta(\theta_m))A(x)$ with conjugate prior $\pi_0(\theta_m|\nu_0, \mathcal{X}_0) = g(\eta(\theta_m))^{\nu_0} \exp\left(\nu_0 \eta(\theta_m)^T \mathcal{X}_0\right) h(\mathcal{X}_0, \nu_0)$ and variational posterior $\widehat{\pi}_m^{\beta_\text{p}}(\theta_m|\nu_m, \mathcal{X}_m) = g(\eta(\theta_m))^{\nu_m} \exp\left(\nu_m \eta(\theta_m)^T \mathcal{X}_m\right) h(\mathcal{X}_m, \nu_m)$ within the same conjugate family is analytically available iff the following three quantities have closed form:

$$\mathbb{E}_{\widehat{\pi}_m^{\beta_\text{p}}}\left[\eta(\theta_m)\right], \ \mathbb{E}_{\widehat{\pi}_m^{\beta_\text{p}}}\left[\log g(\eta(\theta_m))\right], \ \int A(z)^{1+\beta_\text{p}} \left[h\left(\frac{(1+\beta_\text{p})T(z) + \nu_m \mathcal{X}_m}{1+\beta_\text{p}+\nu_m}, 1+\beta+\nu_m \right) \right]^{-1} dz.$$

The conditions of Theorem 2 are met by many exponential models, e.g. the Normal-Inverse-Gamma, the Exponential-Gamma, and the Gamma-Gamma. For a simulated autoregressive BLR, we assess the quality of $\widehat{\pi}^{\beta_\text{p}}$ following Yao et al. [48], who estimate a difference $\hat{k}$ between $\pi_m^{\beta_\text{p}}$ and $\widehat{\pi}_m^{\beta_\text{p}}$ relative to a posterior expectation. We use this on the posterior predictive, which is an expectation relative to $\pi_m^{\beta_\text{p}}$ and drives the CP detection. Yao et al. [48] rate $\widehat{\pi}_m^{\beta_\text{p}}$ as *close* to $\pi_m^{\beta_\text{p}}$ if $\hat{k} < 0.5$. Figs 3 and 2 **B** show that our approximation lies well below this threshold for choices of β_p decreasing reasonably fast with the dimension. Note that these are exactly the values of β_p one will want to select for inference: As d increases, the magnitude of $f_{m_t}(\boldsymbol{y}_t|\mathcal{F}_{t-1})$ decreases rapidly. Hence, β_p needs to decrease as d increases to prevent the β-D inference from being dominated by the integral in Eq. (5) and disregarding $\boldsymbol{y}_t$ [21]. This is also reflected in our experiments in section 5, for which we initialize $\beta_\text{p} = 0.05$ and $\beta_\text{p} = 0.005$ for $d = 1$ and $d = 29$, respectively. However, as Figs. 3 and 2 **B** illustrate, the approximation is still excellent for values of β_p that are much larger than that.

3.2 Stochastic Variance Reduced Gradient (SVRG) for BOCPD

While highest predictive accuracy within BOCPD is achieved using full optimization of the variational parameters at each of T time periods, this has space and time complexity of $\mathcal{O}(T)$ and $\mathcal{O}(T^2)$. In comparison, Stochastic Gradient Descent (SGD) has space and time complexity of $\mathcal{O}(1)$ and $\mathcal{O}(T)$, but yields a loss in accuracy, substantially so for small run-lengths. In the BOCPD setting, there is an obvious trade-off between accuracy and scalability: Since the posterior predictive distributions $f_{m_t}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_t)$ for all run-lengths r_t drive CP detection, SGD estimates are insufficiently accurate for small run-lengths r_t. On the other hand, once r_t is sufficiently large, the variational parameter estimates only need minor adjustments and computing an optimum is costly.

Recently, a new generation of algorithms interpolating SGD and global optimization have addressed this trade-off. They achieve substantially better convergence rates by anchoring the stochastic gradient to a point near an optimum [22, 9, 35, 18, 29]. We propose a memory-efficient two-stage variation of

Stochastic Variance Reduced Gradient (SVRG) inference for BOCPD

```
Input at time 0: Window & batch sizes $W$, $B^*$, $b^*$; frequency $m$, prior $\boldsymbol{\theta}_0$, #steps $K$, step size $\eta$
                 s.t. $W > B^* > b^*$; and $\sim$ denotes sampling without replacement
for next observation $\boldsymbol{y}_t$ at time $t$ do
    for retained run-lengths $r \in R(t)$ do
        if $\tau_r = 0$ then
            if $r < W$ then
                $\boldsymbol{\theta}_r \leftarrow \boldsymbol{\theta}_r^* \leftarrow$ FullOpt $(\text{ELBO}(\boldsymbol{y}_{t-r:t}))$; $\tau_r \leftarrow m$
            else if $r \geq W$ then
                $\boldsymbol{\theta}_r^* \leftarrow \boldsymbol{\theta}_r$; $\tau_r \leftarrow$ Geom $(B^*/(B^* + b^*))$
            $B \leftarrow \min(B^*, r)$
            $g_r^{\text{anchor}} \leftarrow \frac{1}{B}\sum_{i\in\mathcal{I}} \nabla\text{ELBO}(\boldsymbol{\theta}_r^*, \boldsymbol{y}_{t-i})$, where $\mathcal{I} \sim$ Unif$\{0, \dots, \min(r, W)\}$, $|\mathcal{I}| = B$
        for $j = 1, 2, \dots, K$ do
            $b \leftarrow \min(b^*, r)$ and $\widetilde{\mathcal{I}} \sim$ Unif$\{0, \dots, \min(r, W)\}$ and $|\widetilde{\mathcal{I}}| = b$
            $g_r^{\text{old}} \leftarrow \frac{1}{b}\sum_{i\in\widetilde{\mathcal{I}}} \nabla\text{ELBO}(\boldsymbol{\theta}_r^*, \boldsymbol{y}_{t-i})$,  $g_r^{\text{new}} \leftarrow \frac{1}{b}\sum_{i\in\widetilde{\mathcal{I}}} \nabla\text{ELBO}(\boldsymbol{\theta}_r, \boldsymbol{y}_{t-i})$
            $\boldsymbol{\theta}_r \leftarrow \boldsymbol{\theta}_r + \eta \cdot \left(g_r^{\text{new}} - g_r^{\text{old}} + g_r^{\text{anchor}}\right)$; $\tau_r \leftarrow \tau_r - 1$
    $r \leftarrow r + 1$ for all $r \in R(t)$; $R(t) \leftarrow R(t) \cup \{0\}$
```

these methods tailored to BOCPD. First, the variational parameters are moved close to their global optimum using a variant of [22, 35]. Unlike standard versions, we anchor the gradient estimates to a (local) optimum by calling a convex optimizer FullOpt every m steps for the first W iterations. While our implementation uses Python scipy's L-BFSG-B optimization routine, any convex optimizer could be used for this step. Compared to standard SGD or SVRG, full optimization substantially decreases variance and increases accuracy for small r_t. Second, once $r_t > W$ we do not perform full optimization anymore. Instead, we anchor optimization to the current value as in standard SVRG, by updating the anchor at stochastic time intervals determined by a geometric random variable with success probability $B^*/(B^* + b^*)$. Whether the anchor is based on global optimization or not, the next step consists in sampling $B = \min(r_t, B^*)$ observations without replacement from a window with the $\min(r_t, W)$ most recent observations to initiate the SVRG procedure. Following this, for the next K observations, we incrementally refine the estimates while keeping their variance low using a stochastic-batch variant of [29, 30] by sampling a batch of size $b = \min(r_t, b^*)$ without replacement from the $\min(r_t, W)$ most recent observations. The resulting on-line inference has constant space and linear time complexity like SGD, but produces good estimates for small r_t and converges faster [22, 29, 30]. We provide a detailed complexity analysis of the procedure in the Appendix, where we also demonstrate numerically that it is orders of magnitude faster than MCMC-based inference.

4 Choice of β

Initializing β_{p}: The β-D has been used in a variety of settings [15, 4, 14, 49], but there is no principled framework for selecting β. We remedy this by minimizing the expected predictive loss with respect to β on-line. As the losses need not be convex in β_{p}, initial values can matter for the optimization. A priori, we pick β_{p} maximizing the β-D influence for a given Mahalanobis Distance (MD) $\boldsymbol{x}^*$ under $\pi(\theta_m)$. As Figure 1 **A** shows, $\beta_{\text{p}} > 0$ induces a point of maximum influence $\text{MD}(\beta_{\text{p}}, \pi_m(\theta_m))$: Points further in the tails are treated as outliers, while points closer to the mode receive similar influence as under the KLD. A Monte Carlo estimate of $\text{MD}(\beta_{\text{p}}, \pi_m(\theta_m))$ is found via $\widehat{\text{MD}}(\beta_{\text{p}}, \pi_m(\theta_m)) = \text{argmax}_{x\in\mathbb{R}_+} \hat{I}(\beta_{\text{p}}, \pi_m(\theta_m))(x)$ [28]. We initialize β_{p} by solving the inverse problem: For x^*, we seek β_{p} such that $\widehat{\text{MD}}(\beta_{\text{p}}, \pi_m(\theta_m)) = x^*$. (The Appendix contains a pictorial illustration of this procedure.) The k-th standard deviation under the prior is a good choice of x^* for low dimensions [see also 12], but not appropriate as delimiter for high density regions even in moderate dimensions d. Thus, we propose $x^* = \sqrt{d}$ for larger values of d, inspired by the fact that under normality, MD $\rightarrow \sqrt{d}$ as $d \rightarrow \infty$ [17]. One then finds β_{p} by approximating the gradient of $\widehat{\text{MD}}(\beta_{\text{p}}, \pi_m(\theta_m))$ with respect to β_{p}. As β_{rlm} does not affect $\pi_m^{\beta_{\text{p}}}$, its initialization matters less and generally, initializing $\beta_{\text{rlm}} \in [0, 1]$ produces reasonable results.

Optimizing $\boldsymbol{\beta}$ on-line: For $\boldsymbol{\beta} = (\beta_{\text{rlm}}, \beta_{\text{p}})$ and prediction $\widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$ of $\boldsymbol{y}_t$ obtained as posterior expectation via Eq. (6), define $\boldsymbol{\varepsilon}_t(\boldsymbol{\beta}) = \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$. For predictive loss $L : \mathbb{R} \to \mathbb{R}_+$, we target $\boldsymbol{\beta}^* = \text{argmin}_{\boldsymbol{\beta}} \{\mathbb{E}(L(\boldsymbol{\varepsilon}_t(\boldsymbol{\beta})))\}$. Replacing expected by empirical loss and deploying SGD, we seek to find the partial derivatives of $\nabla_{\beta} L(\varepsilon_t(\boldsymbol{\beta}))$. Noting that $\nabla_{\boldsymbol{\beta}} L(\varepsilon_t(\boldsymbol{\beta}))) = L'(\varepsilon_t(\boldsymbol{\beta}))) \cdot \nabla_{\boldsymbol{\beta}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$, the issue reduces to finding the partial derivatives $\nabla_{\beta_{\text{rlm}}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$ and $\nabla_{\beta_{\text{p}}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$. Remarkably, $\nabla_{\beta_{\text{rlm}}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$ can be updated sequentially and efficiently by differentiating the recursion in Eq. (2b). The derivation is provided in the Appendix. The gradient $\nabla_{\beta_{\text{p}}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$ on the other hand is not available analytically and thus is approximated numerically. Now, β can be updated on-line via

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - \eta \cdot \begin{bmatrix} \nabla_{\beta_{\text{rlm},t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right) \\ \nabla_{\beta_{\text{p},t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right) \end{bmatrix} \tag{12}$$

In spirit, this procedure resembles existing approaches for model hyperparameter optimization [8]. For robustness, L should be chosen appropriately. In our experiments L is a bounded absolute loss.

5 Results

Next, we illustrate the most important improvements this paper makes to BOCPD. First, we show how robust BOCPD deals with outliers on the well-log data set. Further, we show that standard BOCPD breaks down in the M-open world whilst β-D yields useful inference by analyzing noisy measurements of Nitrogen Oxide (NOX) levels in London. In both experiments, we use the methods in section 4, on-line hyperparameter optimization [8] and pruning for $p(r_t, m_t|\boldsymbol{y}_{1:t})$ [1]. Detailed information is provided in the Appendix. Software and simulation code is available as part of a reproducibility award at `https://github.com/alan-turing-institute/rbocpdms/`.

5.1 Well-log

The well-log data set was first studied in Ruanaidh et al. [39] and has become a benchmark data set for univariate CP detection. However, except in Fearnhead and Rigaill [12] its outliers have been removed before CP detection algorithms are run [e.g. 1, 31, 40]. With $\mathcal{M}$ containing one BLR model of form $y_t = \mu + \varepsilon_t$, Figure 4 shows that robust BOCPD deals with outliers on-line. The maximum of the run-length distribution for standard BOCPD is zero 145 times, so declaring CPs based on the run-length distribution's maximum [see e.g. 41] yields a false discovery rate (FDR) > 90%. This problem persists even with non-parametric, Gaussian Process, models [p. 186, 45]. Even using Maximum A Posteriori (MAP) segmentation [11], standard BOCPD mislabels 8 outliers as CPs, making for a FDR > 40%. In contrast, the segmentation of the β-D version does not mislabel any outliers. Morevoer and in accordance with Thm. 1, its run-length distribution's maximum never drops to zero in response to outliers. Further, a natural byproduct of the robust segmentation is a reduction in squared (absolute) prediction error by 10% (6%) compared to the standard version. The

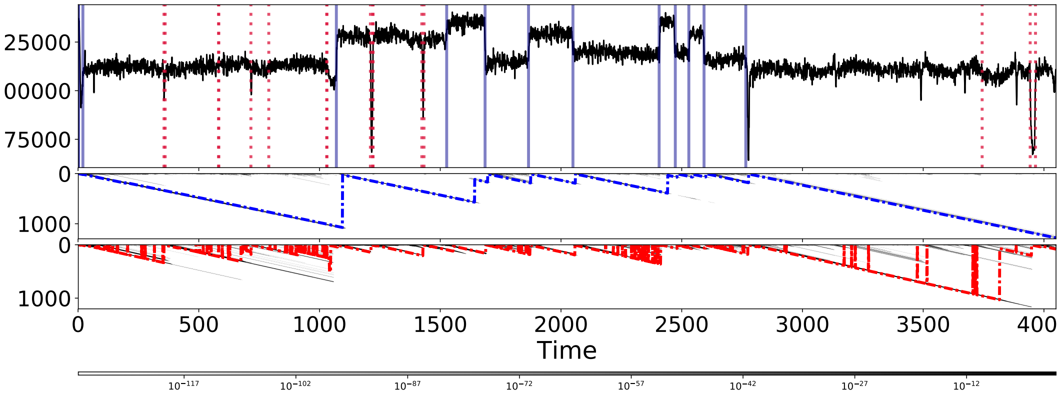

Figure 4: Maximum A Posteriori (MAP) segmentation and run-length distributions of the well-log data. Robust segmentation depicted using solid lines, CPs additionally declared under standard BOCPD with dashed lines. The corresponding run-length distributions for robust (middle) and standard (bottom) BOCPD are shown in grayscale. The most likely run-lengths are dashed.

robust version has more computational overhead than standard BOCPD, but still needs less than 0.5 seconds per observation using a 3.1 GHz Intel i7 and 16GB RAM.

Not only does robust BOCPD's segmentation in Figure 4 match that in Fearnhead and Rigaill [12], but it also offers three additional on-line outputs: Firstly, it produces probabilistic (rather than point) forecasts and parameter inference. Secondly, it self-regulates its robustness via β. Thirdly, it can compare multiple models and produce model posteriors (see section 5.2). Further, unlike Fearnhead and Rigaill [12], it is not restricted to fitting univariate data with piecewise constant functions.

5.2 Air Pollution

The example in Fig. 1 **B** gives an illustration of the importance of robustness in medium-dimensional (BOCPD) problems: It suffices for a *single* dimension of the problem to be misspecified or outlier-prone for inference to fail. Moreover, the presence of misspecification or outliers in this plot can hardly be spotted – and this effect will worsen with increasing dimensionality. To illustrate this point on a multivariate real world data set, we also analyze Nitrogen Oxide (NOX) levels across 29 stations in London using spatially structured Bayesian Vector Autoregressions [see 25]. Previous robust on-line methods [e.g. 36, 7, 12] cannot be applied to this problem because they assume univariate data or do not allow for dependent observations. As Figure 5 shows, robust BOCPD finds one CP corresponding to the introduction of the congestion charge, while standard BOCPD produces an FDR >90%. Both methods find a change in dynamics (i.e. models) after the congestion charge introduction, but variance in the model posterior is substantially lower for the robust algorithm. Further, it increases the average one-step-ahead predictive likelihood by 10% compared to standard BOCPD.

6 Conclusion

This paper has presented the very first robust Bayesian on-line changepoint (CP) detection algorithm and the first ever scalable General Bayesian Inference (GBI) method. While CP detection is a particularly salient example of unaddressed heterogeneity and outliers leading to poor inference, the capabilities of GBI and the Structural Variational approximations presented extend far beyond this setting. With an ever increasing interest in the field of machine learning to efficiently and reliably quantify uncertainty, robust probabilistic inference will only become more relevant. In this paper, we give a particularly striking demonstration of the inferential power that can be unlocked through divergence-based General Bayesian inference.

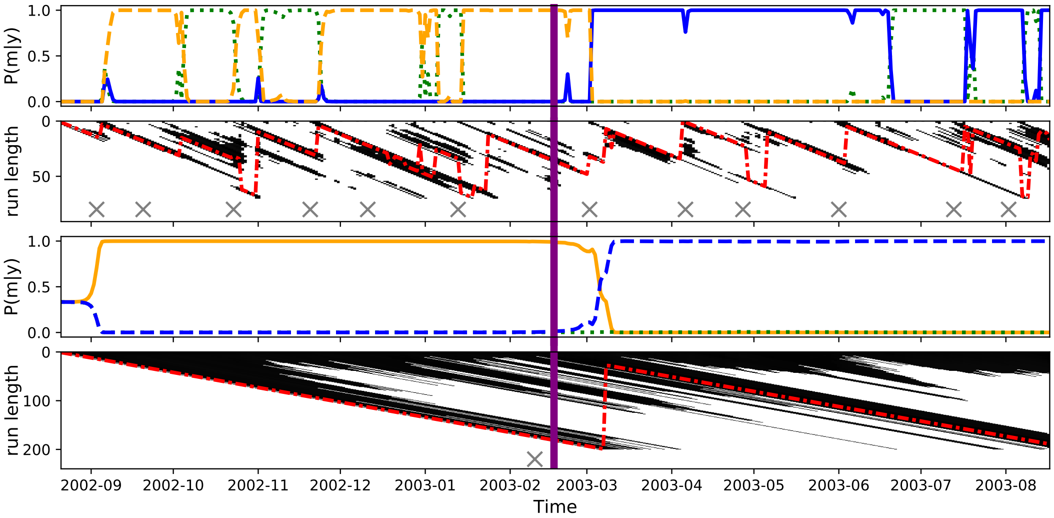

Figure 5: On-line model posteriors for three different VAR models (solid, dashed, dotted) and run-length distributions in grayscale with most likely run-lengths dashed for standard (top two panels) and robust (bottom two panels) BOCPD. Also marked are the congestion charge introduction, 17/02/2003 (solid vertical line) and the MAP segmentations (crosses)

Acknowledgements

We would like to cordially thank both Jim Smith and Chris Holmes for fruitful discussions and help with some of the theoretical results. JK and JJ are funded by EPSRC grant EP/L016710/1 as part of the Oxford-Warwick Statistics Programme (OxWaSP). TD is funded by the Lloyds Register Foundation programme on Data Centric Engineering through the London Air Quality project. This work was supported by The Alan Turing Institute for Data Science and AI under EPSRC grant EP/N510129/1. In collaboration with the Greater London Authority.

References

[1] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[2] Mauricio Alvarez, Jan R Peters, Neil D Lawrence, and Bernhard Schölkopf. Switched latent force models for movement segmentation. In *Advances in neural information processing systems*, pages 55–63, 2010.

[3] Daniel Barry and John A Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

[4] Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[5] José M Bernardo and Adrian FM Smith. Bayesian theory, 2001.

[6] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[7] Yang Cao and Yao Xie. Robust sequential change-point detection by convex optimization. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1287–1291. IEEE, 2017.

[8] François Caron, Arnaud Doucet, and Raphael Gottardo. On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22(2): 579–595, 2012.

[9] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[10] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2729–2738, 2017.

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

[12] Paul Fearnhead and Guillem Rigaill. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, (just-accepted), 2017.

[13] Emily Fox and David B Dunson. Multiresolution Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 737–745, 2012.

[14] Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *Artificial Intelligence and Statistics*, 2018.

[15] Abhik Ghosh and Ayanendranath Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

[16] Marco Grzegorczyk and Dirk Husmeier. Non-stationary continuous dynamic Bayesian networks. In *Advances in Neural Information Processing Systems*, pages 682–690, 2009.

[17] Peter Hall, JS Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

[18] Reza Harikandeh, Mohamed Osama Ahmed, Alim Virani, Mark Schmidt, Jakub Konečnỳ, and Scott Sallinen. Stopwasting my gradients: Practical svrg. In *Advances in Neural Information Processing Systems*, pages 2251–2259, 2015.

[19] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang D Bui, and Richard E Turner. Black-box α-divergence minimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1511–1520. JMLR. org, 2016.

[20] He Huang and Martin Paulus. Learning under uncertainty: a comparison between rw and Bayesian approach. In *Advances in Neural Information Processing Systems*, pages 2730–2738, 2016.

[21] Jack Jewson, Jim Smith, and Chris Holmes. Principles of bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

[22] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[23] Azadeh Khaleghi and Daniil Ryabko. Locating changes in highly dependent data with unknown number of change points. In *Advances in Neural Information Processing Systems*, pages 3086–3094, 2012.

[24] Rebecca Killick, Idris A Eckley, Kevin Ewans, and Philip Jonathan. Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37(13): 1120–1126, 2010.

[25] Jeremias Knoblauch and Theodoros Damoulas. Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2018.

[26] George Konidaris, Scott Kuindersma, Roderic Grupen, and Andrew G Barto. Constructing skill trees for reinforcement learning agents from demonstration trajectories. In *Advances in neural information processing systems*, pages 1162–1170, 2010.

[27] Erich Kummerfeld and David Danks. Tracking time-varying graphical structure. In *Advances in neural information processing systems*, pages 1205–1213, 2013.

[28] Sebastian Kurtek and Karthik Bharath. Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika*, 102(3):601–616, 2015.

[29] Lihua Lei and Michael Jordan. Less than a single pass: Stochastically controlled stochastic gradient. In *Artificial Intelligence and Statistics*, pages 148–156, 2017.

[30] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

[31] Céline Levy-leduc and Zaïd Harchaoui. Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, pages 617–624, 2008.

[32] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

[33] Kevin Lin, James L Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused Lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6887–6896, 2017.

[34] Scott Niekum, Sarah Osentoski, Christopher G Atkeson, and Andrew G Barto. CHAMP: Changepoint detection using approximate model parameters. Technical report, (No. CMU-RI-TR-14-10) Carnegie-Mellon University Pittsburgh PA Robotics Institute, 2014.

[35] Atsushi Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.

[36] Moshe Pollak. A robust changepoint detection method. *Sequential Analysis*, 29(2):146–161, 2010.

[37] Aleksey S Polunchenko, Alexander G Tartakovsky, and Nitis Mukhopadhyay. Nearly optimal change-point detection with an application to cybersecurity. *Sequential Analysis*, 31(3):409–435, 2012.

[38] Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

[39] Ó Ruanaidh, JK Joseph, and William J Fitzgerald. Numerical Bayesian methods applied to signal processing. 1996.

[40] Eric Ruggieri and Marcus Antonellis. An exact approach to Bayesian sequential change point detection. *Computational Statistics & Data Analysis*, 97:71–86, 2016.

[41] Yunus Saatçi, Ryan D Turner, and Carl E Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934, 2010.

[42] Florian Stimberg, Manfred Opper, Guido Sanguinetti, and Andreas Ruttor. Inference in continuous-time change-point models. In *Advances in Neural Information Processing Systems*, pages 2717–2725, 2011.

[43] Ryan Turner, Yunus Saatci, and Carl Edward Rasmussen. Adaptive sequential Bayesian change point detection. In *Temporal Segmentation Workshop at NIPS*, 2009.

[44] Ryan D Turner, Steven Bottone, and Clay J Stanek. Online variational approximations to non-exponential family change point models: with application to radar tracking. In *Advances in Neural Information Processing Systems*, pages 306–314, 2013.

[45] Ryan Darby Turner. *Gaussian processes for state space models and change point detection*. PhD thesis, University of Cambridge, 2012.

[46] Robert C Wilson, Matthew R Nassar, and Joshua I Gold. Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476, 2010.

[47] Xiang Xuan and Kevin Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th international conference on Machine learning*, pages 1055–1062. ACM, 2007.

[48] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538*, 2018.

[49] Kenan Y Yılmaz, Ali T Cemgil, and Umut Simsekli. Generalised coupled tensor factorisation. In *Advances in neural information processing systems*, pages 2151–2159, 2011.

[50] XianXing Zhang, Lawrence Carin, and David B Dunson. Hierarchical topic modeling for analysis of time-evolving personal choices. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2011.