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

Jeremias Knoblauch
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

Theo Damoulas
Department of Statistics
Department of Computer Science
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 predictive 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. We offer the state of the art and improve the False Discovery Rate of CPs by more than 80% on real world data.

1 Introduction

Modeling non-stationary time series with changepoints (CPs) is popular [24, 52, 34] and important in a wide variety of research fields , including genetics [8, 17, 43], finance [28], oceanography [25], brain imaging and cognition [14, 21], cybersecurity [38] and robotics [2, 27]. For streaming data, a particularly important subclass are Bayesian On-line Changepoint Detection (BOCPD) methods that can process data sequentially [1, 12, 44, 49, 48, 42, 8, 35, 45, 41, 26] while providing fullly 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 a strictly increasing influence function. We remedy this by instead minimizing the β-divergence (β-D) whose influence function allows us to deal with outliers effectively, see Fig. 1 **A**. In addressing misspecification and outliers this way, our approach builds on the principles of General

Preprint. Work in progress.

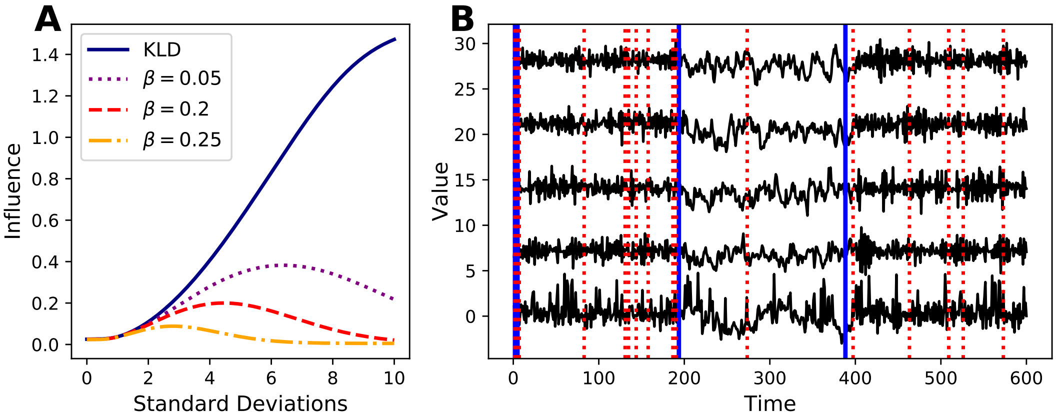

Figure 1: **A:** Influence of $\boldsymbol{y}_t$ on inference as function of distance to the posterior expectation in Standard Deviations for β-divergences with different βs. **B:** Five jointly modeled Simulated Autoregressions (ARs) with true CPs at $t = 200, 400$; bottom-most AR injected with t_4-noise.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

Bayesian Inference (GBI) [see 6, 22] and robust divergences [e.g. 4, 16]. 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 22, 16], 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 [12]. Recently, both formulations have been unified in Knoblauch and Damoulas [26]. The underlying algorithm has extensions ranging from Gaussian Processes [42] and on-line hyperparameter optimization [8] to non-exponential families [45, 35].

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 $q(m_t|m_{t-1}, r_t) = m_{t-1}$ 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 $\boldsymbol{X}_t$ 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 $\mathcal{F}_{t-1} = \left\{\boldsymbol{y}_{1:(t-1)}, r_{t-1}\right\}$ 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 [1, 26].

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 [47, 6]. While this is the most efficient way of updating posterior beliefs if they coincide, this is no longer the case in the M-open world [5] where they match only approximately [22], e.g. in the presence of outliers. GBI [6, 22] 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

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

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

$$\ell^{\text{KLD}}(\boldsymbol{\theta}_m|\boldsymbol{y}_t) = -\log\left(f_{m_t}(\boldsymbol{y}_t|\mathcal{F}_{t-1})\right) \tag{4}$$

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

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. Conversely, under the log score of KLD, *more* influence is associated with an observation the further out in the tails of the model it occurs. 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 $\boldsymbol{y}_t$ [29].

Other divergences than the β-D such as α-Divergences [e.g. 20] also accommodate robust inference. In this work, we restrict ourselves to the β-D as it is the only proper robust divergence not requiring estimation of the DGM's density [22]. 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 [15, 33, 39, 11], but instead framing the procedure as alternative to Variational Bayes.

Apart from the computational gains of section 3.1, we tackle robust inference via the β-D rather than via Student-t errors for three reasons: Firstly, robust run-length posteriors need robustness in *ratios* rather than *tails* (see section 2.3). Secondly, Student-t errors model outliers as part of the DGM, which compromises 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-t under the KLD lets ν_t dominate parameter inference and lets 1% of observations inflate the predictive variance substantially. Thirdly, unlike using Student-t errors, inference with the β-D is applicable to *any* underlying predictive model.

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. 37, 7, 13]. For example, the method in Fearnhead and

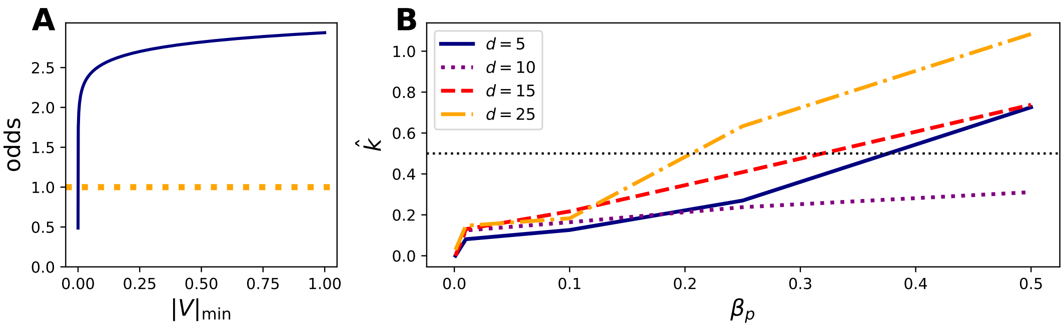

Figure 2: **A**: Lower bound on the odds of Thm. 1 for priors used for Figure 1 **B** and $h(r) = 1/100$. **B**: $\hat{k}$ for different choices of β_{p} and output (input) dimensions d ($2d$) in an autoregressive BLR.

Rigaill [13] 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 $\pi_m(\boldsymbol{\theta}|\boldsymbol{y}_{1:t})$. Thus, we make BOCPD robust by using β-D posteriors $p^{\beta_{\text{rlm}}}(r_t, m_t|\boldsymbol{y}_{1:t})$, $\pi_m^{\beta_{\text{p}}}(\boldsymbol{\theta}|\boldsymbol{y}_{1:t})$ 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 $\exp\{\ell^{\beta_{\text{rlm}}}(\boldsymbol{\theta}_{m_t}|\boldsymbol{y}_t)\}$ of Eq. (5). While $p^{\beta_{\text{rlm}}}(\boldsymbol{y}_{1:t}, r_t, m_t)$ does not integrate to one, $p^{\beta_{\text{rlm}}}(r_t, m_t|\boldsymbol{y}_{1:t})$ still sums to one. 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 [13] is robust because hyperparameters enforce that a single outlier is insufficient for declaring a CP. Analogously, we can quantify robustness by conditioning on $r_t = r$ and studying the odds of $r_{t+1} \in \{0, r+1\}$:

$$\frac{p(r_{t+1} = r+1|\boldsymbol{y}_{1:t+1}, r_t = r, m_t)}{p(r_{t+1} = 0|\boldsymbol{y}_{1:t+1}, r_t = r, m_t)} = \frac{\cancel{p(\boldsymbol{y}_{1:t}, r_t = r, m_t)} \cdot (1 - H(r_{t+1}, r_t)) f_{m_t}^D(\boldsymbol{y}_{t+1}|\mathcal{F}_t)}{\cancel{p(\boldsymbol{y}_{1:t}, r_t = r, m_t)} \cdot H(r_{t+1}, r_t) f_{m_t}^D(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0)}. \tag{7}$$

Here, $f_{m_t}^D$ denotes the negative exponential of the score under divergence D. In particular, $f_{m_t}^{\text{KLD}}(\boldsymbol{y}_{t+1}|\mathcal{F}_t) = f_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$ and $f_{m_t}^{\beta_{\text{rlm}}}(\boldsymbol{y}_{t+1}|\mathcal{F}_t) = \exp\left\{-\ell^{\beta_{\text{rlm}}}(\boldsymbol{\theta}_m|\boldsymbol{y}_t)\right\}$ as in Eq. (5). Taking a closer look at Eq. (7), if $\boldsymbol{y}_{t+1}$ is an outlier with low density under $f_{m_t}^D(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$, the odds will move in favor of a CP provided that the prior is sufficiently uninformative to make $f_{m_t}^D(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0) > f_{m_t}^D(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$. In fact, even very small differences have a substantial impact on the odds. For BLR, Theorem 1 provides conditions guaranteeing that these odds never favor a CP after a single observation under the β-D when they would under the KLD, i.e. when $f_{m_t}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0)$ is much larger than $f_{m_t}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)$.

Theorem 1. *If m_t in Eq. (7) 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 $|V|_{\text{min}} > 0$, then one can choose any $(\beta_{\text{rlm}}, H(r_t, r_{t+1})) \in S\left(p, \beta_{\text{rlm}}, a_0, b_0, \mu_0, \Sigma_0, |V|_{\text{min}}\right)$ to guarantee that*

$$\frac{(1 - H(r_{t+1}, r_t)) f_{m_t}^{\beta_{\text{rlm}}}(\boldsymbol{y}_{t+1}|\mathcal{F}_t)}{H(r_{t+1}, r_t) f_{m_t}^{\beta_{\text{rlm}}}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0)} \geq 1, \tag{8}$$

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

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

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 $|V|_{\min}$.

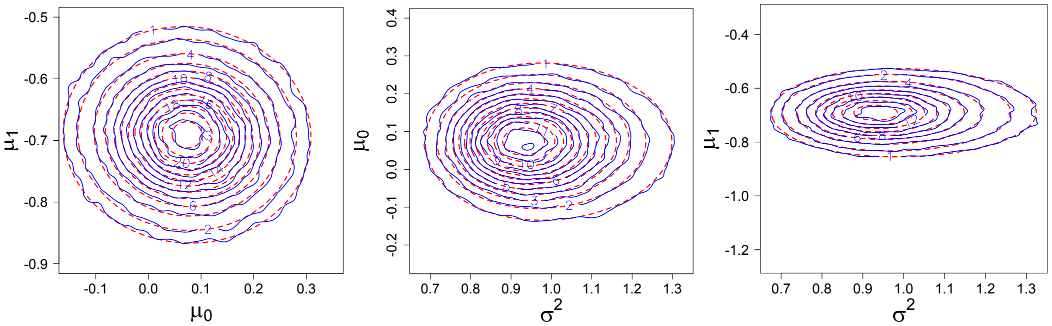

Figure 3: Exemplary contour plots of bivariate marginals for the approximation $\widehat{\pi}_m^{\beta_\text{p}}(\boldsymbol{\theta}_m)$ of Eq. (10) (dashed) and the target $\pi_m^{\beta_\text{p}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t})$ (solid) estimated and smoothed from $95,000$ Hamiltonian 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 Approximation based on pseudo-conjugacy

While there has been a recent surge in theoretical work on GBI [6, 16, 22, 15] applications have been sparse, in large part due to intractability. While MCMC methods have been used successfully for GBI [22, 16], it is hard to scale them for the BOCPD setting: One would have to sample from the parameter posteriors for each run-length and additionally require a second layer of sampling to evaluate the integral in Eq. (5). Circumventing MCMC, most work on BOCPD has focused on conjugate distributions [1, 44, 12] and approximations [45, 35]. 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. Further, since β-D $\to$ KLD as $\beta \to 0$ [4], there is an especially compelling way of doing SVI based on the fact that

$$\pi_m^{\beta_\text{p}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \approx \pi_m^{\text{KLD}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \tag{9}$$

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

$$\widehat{\pi}_m^{\beta_\text{p}}(\boldsymbol{\theta}_m) = \underset{\pi_m^{\text{KLD}}(\boldsymbol{\theta}_m)}{\text{argmin}} \left\{ \text{KL}\left(\pi_m^{\text{KLD}}(\boldsymbol{\theta}_m) \middle\| \pi_m^{\beta_\text{p}}(\boldsymbol{\theta}_m|\boldsymbol{y}_{(t-r_t):t}) \right) \right\}. \tag{10}$$

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 of the KLD-posterior. Further, for many models, optima of the optimization in Eq. (10) can be computed efficiently due to the closed form of its Evidence Lower Bound (ELBO). We state this in Theorem 2 whose proof is in the Appendix, together with the derivation of the ELBO for Bayesian Linear Regression (BLR).

Theorem 2. The ELBO objective corresponding to the β-D posterior approximation in Eq. (10) 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. [50], 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. [50] 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 $\boldsymbol{y}_t$ [22]. 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.

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 $\eta$
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 $(B/(B+b))$
                $g_r^{\text{anchor}} \leftarrow \frac{1}{B}\sum_{i\in\mathcal{I}} \nabla\text{ELBO}(\boldsymbol{\theta}_r^*, \boldsymbol{y}_{t-i})$, where $\mathcal{I} \sim \text{Unif}\{0,\ldots,\min(r,W)\}$, $|\mathcal{I}| = B$
        for $i = 1, 2, \ldots, K$ do
            $\widetilde{\mathcal{I}} \sim \text{Unif}\{0,\ldots,\min(r,W)\}$ 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 $r \in R(t)$; $R(t) \leftarrow R(t) \cup \{0\}$
```

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 [23, 10, 36, 19, 30]. We propose a memory-efficient two-stage variation of these methods tailored to BOCPD. First, the variational parameters are moved close to their global optimum using a variant of [23, 36]. Unlike standard versions, we anchor the gradient estimates to an optimum every m steps for the first W iterations. Compared to standard SGD or SVRG, this substantially decreases variance and increases accuracy for small r_t. Second, once $r_t > W$ we incrementally refine the estimates while keeping their variance low using a stochastic-batch variant of SVRG [30, 31] on a window with the 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 [23, 30, 31]. We provide a detailed complexity analysis of the procedure in the Appendix Compared to MCMC-based inference, it is orders of magnitude faster: E.g. for the well-log data in section 5.1, an MCMC implementation in `Stan` [9] takes 10^5 times longer.

4 Choice of β

Initializing β_p: The β-D has been used in a variety of settings [16, 4, 15, 51], but there is no principled framework for selecting $\boldsymbol{\beta}$. 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 $\widehat{\text{MD}}(\beta_{\text{p}}, \pi_m(\theta_m)) = \text{argmax}_{x \in \mathbb{R}_+} \hat{I}(\beta_{\text{p}}, \pi_m(\theta_m))(x)$ [29]. 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 k-th standard deviation under the prior is a good choice of x^* for low dimensions [see also 13], 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, $\text{MD} \to \sqrt{d}$ as $d \to \infty$ [18]. 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 $\boldsymbol{\beta}^* = \text{argmin}_{\boldsymbol{\beta}} \{\mathbb{E}\left(L(\boldsymbol{\varepsilon}_t(\boldsymbol{\beta}))\right)\}$. Replacing expected by empirical loss and deploying SGD, we seek to find the partial derivatives of $\nabla_{\beta} L\left(\varepsilon_t(\boldsymbol{\beta})\right)$. Noting that $\nabla_{\boldsymbol{\beta}} L\left(\varepsilon_t(\boldsymbol{\beta})\right)) = L'\left(\varepsilon_t(\boldsymbol{\beta})\right)) \cdot \nabla_{\boldsymbol{\beta}}\, \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$, 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

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - \eta \cdot \begin{bmatrix} \nabla_{\beta_{\text{rlm},t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right)\\ \nabla_{\beta_{\text{p},t}} L\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right)\end{bmatrix} \tag{11}$$

In spirit, this procedure resembles existing approaches for model hyperparameter optimization [8]. For robustness, L should be chosen appropriately. Thus, in our experiments we use $L(x) = |x|$.

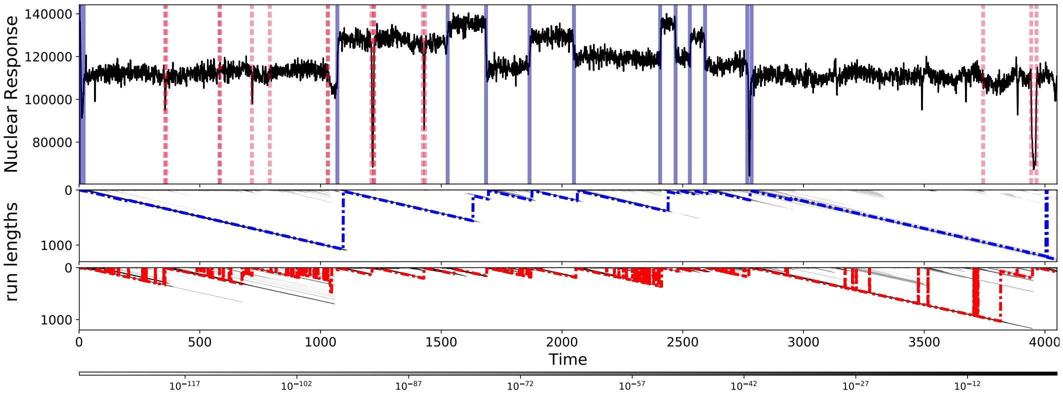

Figure 4: Maximum A Posteriori (MAP) segmentation and run-length distributions of the well-log data. Robust segmentation depicted using solid lines, CPs additionally declared under standard BOCPD with dashed lines. The corresponding run-length distributions for robust (middle) and standard (bottom) BOCPD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

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 are available at XXXXX.

5.1 Well-log

The well-log data set was first studied in Ruanaidh et al. [40] and has become a benchmark data set for univariate CP detection. However, except in Fearnhead and Rigaill [13] its outliers have been removed before CP detection algorithms are run [e.g. 1, 32, 41]. 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. 42] yields a false discovery rate (FDR) > 90%. This problem persists even with non-parametric, Gaussian Process, models [p. 186, 46]. Even using Maximum A Posteriori (MAP) segmentation [12], 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. Further and in accordance with Thm. 1, its run-length distribution's maximum falsely drops to a zero run-length only once, which is in response to >20 consecutive outliers. A natural byproduct of the robust segmentation is a reduction in mean square (absolute) prediction error by 10% (6%) compared to the standard version. The 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 [13], 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 $\boldsymbol{\beta}$. Thirdly, it can compare multiple models and produce model posteriors (see section 5.2). Further, unlike Fearnhead and Rigaill [13], it is not restricted to fitting univariate data with piecewise constant functions.

5.2 Air Pollution

We apply robust BOCPD to analyze Nitrogen Oxide (NOX) levels across 29 stations in London using spatially structured Bayesian Vector Autoregressions (VARs) [see 26]. Previous robust on-line methods [e.g. 37, 7, 13] 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.

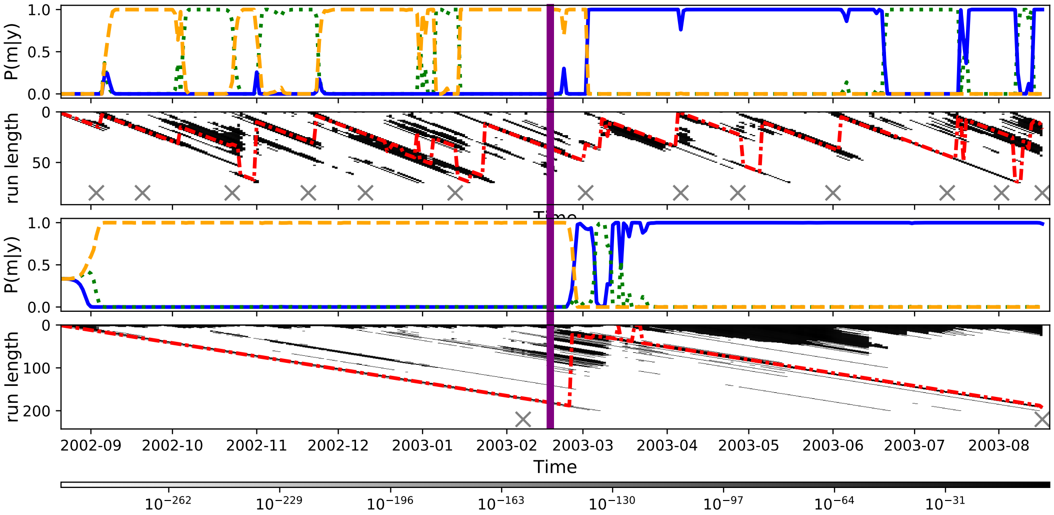

Figure 5: On-line model posteriors for three different VAR models (solid, dashed, dotted) and run-length distributions in grayscale with most likely run-lengths dashed for standard (top two panels) and robust (bottom two panels) BOCPD. Also marked are the congestion charge introduction, 17/02/2003 (solid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

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.

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] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[10] 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.

[11] 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.

[12] 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.

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

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

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Jack Jewson, Jim Q Smith, and Chris Holmes. Principled Bayesian minimum divergence inference. *arXiv preprint arXiv:1802.09411*, 2018.

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

[24] 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.

[25] 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.

[26] 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-18)*, 2018. to appear.

[27] 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.

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

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

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

[31] 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.

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

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

[34] 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.

[35] 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.

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

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

[38] 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.

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

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

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

[42] 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.

[43] 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.

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

[45] 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.

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

[47] Stephen G Walker. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10):1621–1633, 2013.

[48] 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.

[49] 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.

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

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

[52] 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.