Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with $\beta$-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 $\beta$-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 $\beta$ 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 [11, 12, 44, 49, 48, 28, 35, 45, 41, 26] while providing fully probabilistic uncertainty quantification. These algorithms declare CPs if the posterior predictive computed from $y_{1:t}$ at time $t$ has low density for the value of the observation $y_{t+1}$ 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 $\beta$-divergence ($\beta$-D) whose influence function allows us to deal with outliers effectively, see Fig. 1A. In addressing misspecification and outliers this way, our approach builds on the principles of General
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 discriminates outliers and CPs, see Fig. 1B.

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 \( \beta \)**: While Fig. 1A shows that \( \beta \) 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 \( \beta \). 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 \( \beta \)-D. We quantify the degree of robustness and show that inference under the \( \beta \)-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 \( \beta \)-D posterior. Within BOCPD, we propose to scale SVI using variance-reduced Stochastic Gradient Descent. Next, Section 4 expands on how \( \beta \) 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 \( \beta \)-Divergences

BOCPD is based on the Product Partition Model (3) and introduced independently in Adams and MacKay [11] 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 \( \{ y_t \}_{t=1}^{\infty} \) 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

\[
\begin{align*}
    r_t | r_{t-1} & \sim H(r_t, r_{t-1}) & m_t | m_{t-1}, r_t & \sim q(m_t | m_{t-1}, r_t) \\
\theta_{m_t} | m_t & \sim \pi_{m_t} (\theta_{m_t}) & y_t | m_t, \theta_{m_t} & \sim f_{m_t} (y_t | \theta_{m_t})
\end{align*}
\]  

(1a) (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...
Apart from the computational gains of section 3.1, we tackle robust inference via the \( \beta \) Variational Bayes. A where the run-length and model posteriors are then available exactly at time \( t \) for all models \( m \in M \), so then is the recursive computation given by

\[
p(y_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \{ f_{m_t}(y_{1:t}|F_{t-1}) q(m_t|F_{t-1}, m_{t-1}) H(r_t, r_{t-1}) p(y_{1:t-1}, r_{t-1}, m_{t-1}) \} \tag{2b}
\]

where \( F_{t-1} = \{ y_{1:t-1}, r_{t-1} \} \) and \( p(y_{1:t}, r_t, m_t) \) is the joint density of \( y_{1:t}, m_t \) and \( r_t \). The run-length and model posteriors are then available exactly at time \( t \), as \( p(r_t, m_t|y_{1:t}) = p(y_{1:t-1}, r_{t-1}, m_t) = p(y_{1:t}, r_t, m_t) / \sum_{m_t, r_t} p(y_{1:t}, r_t, m_t) \). For a full derivation and the resulting inference see [11] [26].

2.1 General Bayesian Inference (GBI) with \( \beta \)-Divergences (\( \beta \)-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 \( \ell \) and divergences \( D \) to deduce for \( D \) a corresponding loss \( \ell^D \). It can then be shown that for model \( m_t \), the posterior update optimal for \( D \) yields the distribution

\[
\pi^D_{m_t}(\theta_m|y_{1:t-1}; t) \propto \pi_m(\theta) \exp \left\{ -\sum_{i=t-r_t}^{t} \ell^D(\theta_m|y_i) \right\}. \tag{3}
\]

For the KLD and \( \beta \)-D, these losses are the log score and the Tsallis score:

\[
\ell^{\text{KLD}}(\theta_m|y_i) = -\log(f_{m_t}(y_{1:t}|F_{t-1})) \tag{4}
\]

\[
\ell^{\beta}(\theta_m|y_i) = -\left( \frac{1}{\beta_p} f_{m_t}(y_{1:t}|F_{t-1})^{\beta_p} - \frac{1}{1+\beta_p} \right) \int_{y} f_{m_t}(z|F_{t-1})^{1+\beta_p} dz. \tag{5}
\]

Eq. (5) shows why the \( \beta \)-D excels at robust inference: Similar to tempering, \( \ell^\beta \) 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 \( I(y_{1:t}) \) in Figure 1A. I (\( y_i \)) is a divergence between the posterior with and without an observation \( y_i \) [29].

Other divergences than the \( \beta \)-D such as \( \alpha \)-Divergences [e.g. 20] also accommodate robust inference. In this work, we restrict ourselves to the \( \beta \)-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] [53] [3], 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 \( \beta \)-D rather than via Student-\( t \) errors for three reasons: Firstly, robust run-length posteriors need robustness in ratios rather than tails (see section 3.1). Secondly, Student-\( t \) errors model outliers as part of the DGM, which compromises the inference target: Consider a BLR with error \( e_t = \xi_t + \nu_t \nu_{\nu} \), where \( \nu_t \sim \text{Ber}(p) \) for \( p = 0.01 \), \( \nu_t \sim N(0, \sigma^2) \) with outliers \( \nu_{\nu} \sim t_3(0, \gamma) \). Appropriate choices of \( \beta_p \) 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 \( \nu_{\nu} \) dominate parameter inference and lets 1\% of observations inflate the predictive variance substantially. Thirdly, unlike using Student-\( t \) errors, inference with the \( \beta \)-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. 57] [7] [13]. For example, the method in Fearnhead and
Rigall [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

$$E\left( y_t | y_{1:(t-1)} \right) = \sum_{r_t, m_t} E\left( y_t | y_{1:(t-1)}, r_{t-1}, m_{t-1} \right) p(r_{t-1}, m_{t-1} | y_{1:(t-1)}) \tag{6}$$

the key observation is that prediction is driven by two probability distributions: The run-length and model posterior $p(r_t, m_t | y_{1:t})$ and parameter posterior distributions $\pi_m(\theta | y_{1:t})$. Thus, we make BOCPD robust by using $\beta$-D posteriors $p^{\beta_m}(r_t, m_t | y_{1:t})$, $\pi_m^\beta(\theta | y_{1:t})$ for $\beta = (\beta_{\text{hm}}, \beta_p)$.

$\beta_{\text{hm}}$ prevents abrupt changes in $p^{\beta_m}(r_t, m_t | y_{1:t})$ 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 $f_m(y_t | y_m)$ and $f_m(y_t | F_{t-1})$ by their $\beta$-D-counterparts $\exp\{\ell_{\beta_m}(\theta, y_t)\}$ of Eq. (5). While $p^{\beta_m}(y_{1:t}, r_t, m_t)$ does not integrate to one, $p^{\beta_m}(r_t, m_t | y_{1:t})$ still sums to one. Complementing this, $\beta_p$ regulates the robustness of $\pi_m^\beta(\theta | y_{1:t})$ by preventing it from being dominated by tail events. Section 3.1 overcomes the intractability of $\pi_m^\beta(\theta | y_{1:t})$ using Structural Variational Inference (SVI) that recovers the approximated distribution exactly as $\beta_p \to 0$.

### 2.3 Quantifying robustness

The algorithm of Fearnhead and Rigall [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 $r_{t+1} \in \{0, r + 1\}$:

$$\frac{p(r_{t+1} = r+1 | y_{1:t+1}, r_t = r, m_t)}{p(r_{t+1} = 0 | y_{1:t+1}, r_t = r, m_t)} = \frac{p(y_{t+1}, r=r, m_t) \cdot (1 - H(r_{t+1}, r_t)) f_m(y_{t+1} | F_t)}{p(y_{t+1}, r=r, m_t) \cdot H(r_{t+1}, r_t) f_m^D(y_{t+1} | y_0)} \tag{7}$$

Here, $f_m^D$ denotes the negative exponential of the score under divergence $D$. In particular, $f_m^D(y_{t+1} | F_t) = f_m(y_{t+1} | F_t)$ and $f_m^D(y_{t+1} | y_0) = \exp\{-\ell_{\beta_m}(\theta, y_t)\}$ as in Eq. (6). Taking a closer look at Eq. (7), if $y_{t+1}$ is an outlier with low density under $f_m^D(y_{t+1} | F_t)$, the odds will move in favor of the prior is sufficiently uninformative to make $f_m^D(y_{t+1} | y_0) > f_m^D(y_{t+1} | F_t)$. 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 $\beta$-D when they would under the KLD, i.e. when $f_m(y_{t+1} | y_0)$ is much larger than $f_m(y_{t+1} | F_t)$.

**Theorem 1.** If $m_t$ in Eq. (7) is the Bayesian Linear Regression (BLR) model with $\mu \in \mathbb{R}^d$ and priors $a_0, b_0, \mu_0, \Sigma_0$; and if the posterior predictive’s variance determinant is larger than $|V_{\text{min}}| > 0$, then one can choose any $(\beta_{\text{hm}}, H(r_{t+1}, r_t)) \in S(p, \beta_{\text{hm}}, a_0, b_0, \mu_0, \Sigma_0, |V_{\text{min}}|)$ to guarantee that

$$\frac{(1 - H(r_{t+1}, r_t)) f_m^D(y_{t+1} | F_t)}{H(r_{t+1}, r_t) f_m^D(y_{t+1} | y_0)} \geq 1,$$

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

1In fact, $\beta_p = \beta_{\text{hm}}^m$, i.e. the robustness is model-specific, but this is suppressed for readability.
While this ensures that the densities $\pi^\beta_m(\theta_m)$ of Eq. (10) (dashed) and the target $\hat{\pi}^\beta_m(\theta_m|y_{(t-r_t):t})$ (solid) estimated and smoothed from 95,000 Hamiltonian Monte Carlo samples for the $\beta$-D posterior of BLR with $d = 1$, two regressors and $\beta_p = 0.25$.

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 $\beta$-D $\rightarrow$ KLD as $\beta \rightarrow 0$ [4], there is an especially compelling way of doing SVI based on the fact that

$$\hat{\pi}^\beta_m(\theta_m) = \arg\min_{\pi^{KLD}_{m}(\theta_m)} \left\{ K L \left( \hat{\pi}^\beta_m(\theta_m) \Vdash \pi^{KLD}_{m}(\theta_m) \right) \right\}.$$  \hfill (10)

While this ensures that the densities $\hat{\pi}^\beta_m$ and $\pi^{KLD}_{m}$ 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 $\beta$-D posterior approximation in Eq. (10) of an exponential family likelihood model $f_m(y; \theta_m) = \exp \left( \eta(\theta_m)^T T(y) \right) g(\eta(\theta_m)) A(x)$ with conjugate prior $\pi_0(\theta_m|\nu, \lambda_0) = g(\eta(\theta_m)) \nu_m \exp (\nu_0 \eta(\theta_m)^T X_0) h(x, \nu_0)$ and variational posterior $\hat{\pi}^\beta_m(\theta_m|\nu_m, \lambda_m) = g(\eta(\theta_m)) \nu_m \exp (\nu_0 \eta(\theta_m)^T X_0) h(x, \lambda_m)$ within the same conjugate family is analytically available iff the following three quantities have closed form:

$$\mathbb{E}_{\hat{\pi}^\beta_m}[\eta(\theta_m)], \quad \mathbb{E}_{\hat{\pi}^\beta_m} [\log g(\eta(\theta_m))], \quad \int A(z)^{1+\beta_p} h \left( \frac{(1+\beta_p)T(z) + \nu_m X_m}{1 + \beta_p + \nu_m}, 1 + \beta + \nu_m \right) dz.$$  

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 \( \hat{\pi}^\beta_r \) following Yao et al. [50], who estimate a difference \( \hat{k} \) between \( \pi_m^\beta \) and \( \hat{\pi}_m^\beta \) relative to a posterior expectation. We use this on the posterior predictive, which is an expectation relative to \( \pi_m^\beta \) and drives the CP detection. Yao et al. [50] rate \( \hat{\pi}_m^\beta \) as close to \( \pi_m^\beta \) if \( \hat{k} < 0.5 \). Figs. 3 and 2B show that our approximation lies well below this threshold for choices of \( \beta_p \) decreasing reasonably fast with the dimension. Note that these are exactly the values of \( \beta_p \) one would want to select for inference: As \( d \) increases, the magnitude of \( f_m(y; h(t-1)) \) decreases rapidly. Hence, \( \beta_p \) needs to decrease as \( d \) increases to prevent the \( \beta \)-D inference from being dominated by the integral in Eq. 5 and disregarding \( y \). This is also reflected in our experiments in section 5.1, for which we initialize \( \beta_p = 0.05 \) and \( \beta_p = 0.005 \) for \( d = 1 \) and \( d = 29 \), respectively. However, as Figs. 3 and 2B illustrate, the approximation is still excellent for values of \( \beta_p \) 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 \( O(T) \) and \( O(T^2) \). In comparison, Stochastic Gradient Descent (SGD) has space and time complexity of \( O(1) \) and \( O(T) \), 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 \( f_m(y; h(t)) \) 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.

<table>
<thead>
<tr>
<th>Stochastic Variance Reduced Gradient (SVRG) inference for BOCPD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input at time</strong> 0: Window &amp; batch sizes ( W, B; b ); frequency ( m ), prior ( \theta_0 ), # steps ( K ), step size ( \eta )</td>
</tr>
<tr>
<td><strong>for</strong> next observation ( y_t ) at time ( t ) do</td>
</tr>
<tr>
<td>for retained run-lengths ( r \in R(t) ) do</td>
</tr>
<tr>
<td>if ( \tau_r = 0 ) then</td>
</tr>
<tr>
<td>if ( r &lt; W ) then</td>
</tr>
<tr>
<td>( \theta_r \leftarrow \theta_r^* \leftarrow \text{FullOpt}(\text{ELBO}(y_{t-r:t})) ); ( \tau_r \leftarrow m )</td>
</tr>
<tr>
<td>else if ( r \geq W ) then</td>
</tr>
<tr>
<td>( \theta_r^* \leftarrow \theta_r; \tau_r \leftarrow \text{Geom}(B/(B+b)) )</td>
</tr>
<tr>
<td>( g_r^{\text{anchor}} \leftarrow \frac{1}{m} \sum_{i \in I} \nabla \text{ELBO}(\theta_r^*, y_{t-i}), \text{where } I \sim \text{Unif}[0, \ldots, \min(r, W)],</td>
</tr>
<tr>
<td>for ( i = 1, 2, \ldots, K ) do</td>
</tr>
<tr>
<td>( \bar{I} \sim \text{Unif}[0, \ldots, \min(r, W)] } \text{ and }</td>
</tr>
<tr>
<td>( g_r^{\text{old}} \leftarrow \frac{1}{b} \sum_{i \in \bar{I}} \nabla \text{ELBO}(\theta_r^*, y_{t-i}) )</td>
</tr>
<tr>
<td>( g_r^{\text{new}} \leftarrow \frac{1}{b} \sum_{i \in \bar{I}} \nabla \text{ELBO}(\theta_r, y_{t-i}) )</td>
</tr>
<tr>
<td>( \theta_r \leftarrow \theta_r + \eta \cdot (g_r^{\text{new}} - g_r^{\text{old}} + g_r^{\text{anchor}}); \tau_r \leftarrow \tau_r - 1 )</td>
</tr>
<tr>
<td>( r \leftarrow r + 1 ) for all ( r \in R(t); R(t) \leftarrow R(t) \cup {0} )</td>
</tr>
</tbody>
</table>

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 \( \beta \)

Initializing \( \beta_p \): The \( \beta \)-D has been used in a variety of settings \([16, 4, 15, 51]\), but there is no principled framework for selecting \( \beta \). We remedy this by minimizing the expected predictive loss.
with respect to $\beta$ on-line. As the losses need not be convex in $\beta_p$, initial values can matter for the optimization. A priori, we pick $\beta_p$ maximizing the $\beta$-D influence for a given Mahalanobis Distance (MD) $x^*$ under $\pi(\theta_m)$. As Figure 1 shows, $\beta_p > 0$ induces a point of maximum influence $\text{MD}(\beta_p, \pi_m(\theta_m))$: Points further in the tails are treated as outliers, while points closer to the mode receive similar influence as under the K.L.D. A Monte Carlo estimate of $\text{MD}(\beta_p, \pi_m(\theta_m))$ is found via $\text{MD}(\beta_p, \pi_m(\theta_m)) = \arg\max_{x \in \mathbb{R}} I(\beta_p, \pi_m(\theta_m))(x)$ [29]. We initialize $\beta_p$ by solving the inverse problem: For $x^*$, we seek $\beta_p$ such that $\text{MD}(\beta_p, \pi_m(\theta_m)) = x^*$. 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, $\text{MD} \rightarrow \sqrt{d}$ as $d \rightarrow \infty$ [18]. One then finds $\beta_p$ by approximating the gradient of $\text{MD}(\beta_p, \pi_m(\theta_m))$ with respect to $\beta_p$. As $\beta_{\text{im}}$ does not affect $\pi_m^0$, its initialization matters less and generally, initializing $\beta_{\text{im}} \in [0, 1]$ produces reasonable results.

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

$$\beta_t = \beta_{t-1} - \eta \cdot \left[ \frac{\nabla \beta_{\text{im}} L(\varepsilon_t(\beta_{t-1}))}{\nabla \beta_{\text{im}} L(\varepsilon_t(\beta_{t-1}))} \right]$$  \hspace{0.5cm} (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.](image)

### 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 $\beta$-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 $p(r_t, m_i; \hat{y}_{1:t})$ [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 Rigail [13] its outliers have been removed before CP detection algorithms are run [e.g. [1, 32, 41]. With $M$ containing one BLR model of form $y_t = \mu + \epsilon_t$, 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 $\beta$-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 Rigail [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 $\beta$. Thirdly, it can compare multiple models and produce model posteriors (see section 5.2). Further, unlike Fearnhead and Rigail [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)
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


