Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection

Jeremias Knoblauch [1] **Theodoros Damoulas** [1 2 3]

Abstract

Bayesian On-line Changepoint Detection is extended to on-line model selection and non-stationary spatio-temporal processes. We propose spatially structured Vector Autoregressions (VARs) for modelling the process between changepoints (CPs) and give an upper bound on the approximation error of such models. The resulting algorithm performs prediction, model selection and CP detection on-line. Its time complexity is linear and its space complexity constant, and thus it is two orders of magnitudes faster than its closest competitor. In addition, it outperforms the state of the art for multivariate data.

1. Introduction

Real-world spatio-temporal processes are often poorly modelled by standard inference methods that assume stationarity in time and space. A variety of techniques have been developed for modelling non-stationarity in time via changepoints (CPs), ranging from methods for Gaussian Processes (GPs) (Garnett et al., 2009), the Lasso (Lin et al., 2017) or the Ising model (Fazayeli & Banerjee, 2016) over approaches using density ratio estimation (Liu et al., 2013) and kernel-based methods exploiting M-statistics (Li et al., 2015) to framing CP detection as time series clustering (Khaleghi & Ryabko, 2014). In contrast, CP inference allowing for non-stationarity in space (Herlands et al., 2016) has received comparatively little attention.

We offer the first on-line solution to this problem by modeling non-stationarity in both space and time. CPs are used to model non-stationarity in time, and the use of spatially structured Bayesian Vector Autoregressions (SSBVAR) circumvents the assumption of stationarity in space. We unify the approaches in Adams & MacKay (2007) and Fearnhead & Liu (2007) by building a flexible algorithm with three on-line outputs: *prediction*, *model posteriors* and a *MAP segmentation* for CPs and models. This conjunction is possible as both methods build on the Product Partition Model (Barry & Hartigan, 1993). The inputs for the algorithm are a multivariate data stream $\{\boldsymbol{Y}_t\}$ and a set of Bayesian models $\mathcal{M}$ that describe the data between CPs. The algorithm's advantages over other procedures are threefold: Firstly, it performs better than competing algorithms at a fraction of the computational cost. Secondly, we distinguish between data generating mechanisms with different and interpretable parameterizations via Bayesian model selection. Thirdly, the algorithm lends itself naturally to efficient high-dimensional spatio-temporal inference using spatial neighbourhoods.

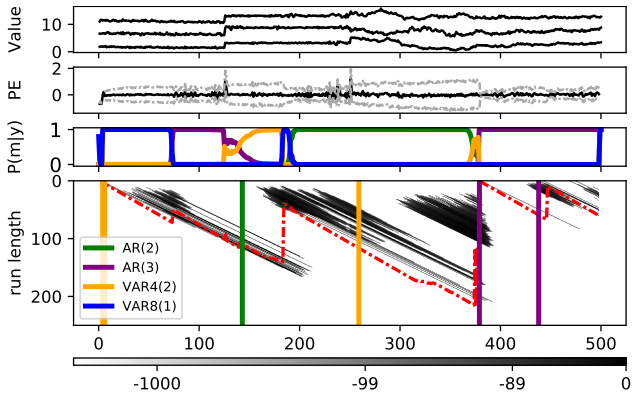

Figure 1. Bayesian On-line Changepoint Detection with Model Selection (BOCPDMS): **Panel 1:** Artificial data across times $1-500$ for a regular spatial grid with 4- and 8-neighbourhood dependency structure as in Fig. 2, **Panel 2:** prediction error (black) and variance (gray). **Panel 3:** Model posteriors $\mathbb{P}(m_t|\boldsymbol{y}_{1:t})$. **Panel 4:** log run-length distribution (grayscale), its maximum (red) and MAP segmentation of CPs and models in corresponding colors.

In spirit, our work is similar to Xuan & Murphy (2007), which performs automatic *off-line* model selection and inference on dependent multivariate time series. In particular, their dependence is modelled using correlated errors in autoregressions. The approach requires conjugate priors to scale, restricting the dependency patterns to decomposable graphs. In contrast, our inference procedure requires specification of a model universe $\mathcal{M}$ before running the algorithm,

[1]Department of Statistics, University of Warwick, UK [2]Department of Computer Science, University of Warwick, UK [3]The Alan Turing Institute for Data Science, UK. Correspondence to: Jeremias Knoblauch <j.knoblauch@warwick.ac.uk>.

To appear in Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

but inference is *on-line*. Crucially, no restrictions are imposed on dependencies in $\mathcal{M}$. This is achieved by modelling dependency *conditionally* rather than *contemporaneously*.

The second line of related research is developed in Saatçi et al. (2010), which extends the Bayesian On-line Changepoint Detection (BOCPD) algorithm of Adams & MacKay (2007) to Gaussian Process (GP) CP models. These are compatible with our framework as elements of the model universe $\mathcal{M}$. However, GPs incur additional computational cost, increasing the overall complexity by two orders of magnitude. The proposed approach avoids this and outperforms the state of the art in the multivariate setting, while offering comparable performance in the univariate one.

The structure of this paper is as follows: Section 2 generalizes the BOCPD algorithm of Adams & MacKay (2007), henceforth AM, by integrating it with the approach of Fearnhead & Liu (2007), henceforth FL. In so doing, we arrive at BOCPD with Model Selection, henceforth BOCPDMS. Section 3 proposes VAR models for non-stationary processes within the BOCPD framework. This motivates populating the model universe $\mathcal{M}$ with spatially structured BVAR (SSBVAR) models. Sections 4–5 address computational aspects. Section 6 demonstrates the algorithm's advantages on real world data.

2. BOCPDMS

Let $\{\boldsymbol{Y}_t\}$ be a data stream with an unknown number of CPs. Focusing on univariate data, FL and AM tackled inference by tracking the posterior distribution for the most recent CP. While FL allow the data to be described by different models between CPs, AM only allow for a single model. However, AM perform 1-step-ahead predictions, whereas FL do not. Instead, they propose a Maximum A Posteriori (MAP) segmentation for CPs and models. In the remainder of this section, we unify both inference approaches. We call the resulting algorithm BOCPD with model selection (BOCPDMS), as it performs prediction, MAP segmentation and model selection on-line.

2.1. Run-length & model universe

The *run-length* r_t at time t is defined as the time since the most recent CP at time t, so $r_t = 0$ corresponds to a CP at time t. Suppose that data between successive CPs can be described by Bayesian models collected in the *model universe* $\mathcal{M}$. For the process $\{\boldsymbol{Y}_t\}$ on $\mathbb{R}^S$, a model $m \in \mathcal{M}$ consist of a conditional probability density $d\mathbb{P}(\boldsymbol{Y}_t|\boldsymbol{\theta}_m)$ on $\mathbb{R}^S$ and a parameter prior density $d\mathbb{P}(\boldsymbol{\theta}_m)$ on $\boldsymbol{\Theta}_m$ depending on hyper-parameters $\boldsymbol{\theta}_m^0$. The notion of $\mathcal{M}$ is due to FL and allows for model uncertainty amongst models developed for BOCPD. For instance, $m \in \mathcal{M}$ could be a GP (Saatçi et al., 2010), a time-deterministic regression (Fearnhead, 2005) or a mixture distribution (Caron et al., 2012).

BOCPD with Model Selection (BOCPDMS)

Input at time 0**:** model universe $\mathcal{M}$; hazard H; prior q
Input at time t**:** next observation $\boldsymbol{y}_t$
Output at time t: $\widehat{\boldsymbol{y}}_{(t+1):(t+h_{\max})}$, S_t, $\mathbb{P}(m_t|\boldsymbol{y}_{1:t})$

for next observation $\boldsymbol{y}_t$ at time t **do**

// STEP I: Compute model-specific quantities
for $m \in \mathcal{M}$ **do**
if $t-1 =$ lag_length(m) **then**
I.A Initialize $d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = 0, m_t = m)$ with prior
else if $t-1 >$ lag_length(m) **then**
I.B.1 Update $d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t, m_t = m)$ via (5a), (5b)
I.B.2 Prune model-specific run-length distribution
I.B.3 Perform hyperparameter inference via (12)
end if
end for

// STEP II: Aggregate over models
if $t >= \min(\text{lag\_length(m)})$ **then**
II.1 Obtain joint distribution over $\mathcal{M}$ via (6a)–(6f)
II.2 Compute (7)–(9)
II.3 **Output:** $\widehat{\boldsymbol{y}}_{(t+1):(t+h_{\max})}, S_t, \mathbb{P}(m_t|\boldsymbol{y}_{1:t})$
end if
end for

2.2. Probabilistic formulation & recursions

Denote by m_t the model describing $\boldsymbol{y}_{(t-r_t):t}$, i.e. the data since the last CP. Given hazard function $H : \mathbb{N} \to [0,1]$, and model prior $q : \mathcal{M} \to [0,1]$, the prior belief is

$$\mathbb{P}(r_t|r_{t-1}) = \begin{cases} 1 - H(r_{t-1}+1) & \text{if } r_t > 0 \\ H(r_{t-1}+1) & \text{if } r_t = 0 \\ 0 & \text{otherwise.} \end{cases} \quad (1a)$$

$$\mathbb{P}(m_t|m_{t-1}, r_t) = \begin{cases} \mathbb{1}_{m_{t-1}}(m_t) & \text{if } r_t > 0 \\ q(m_t) & \text{if } r_t = 0. \end{cases} \quad (1b)$$

Eq. (1b) implies that the model at time t will be equal to the model at time $t-1$ unless a CP occured at t, in which case the next model m_t will be a random draw from q. At time t, the algorithm requires for all possible models m_t and run-lengths r_t the computation of the densities

$$d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, m_t, r_t) = \int_{\Theta_{m_t}} d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{\theta}_{m_t}) d\mathbb{P}(\theta_{m_t}|\boldsymbol{y}_{(t-r_t):(t-1)}) d\boldsymbol{\theta}_{m_t}. \quad (2)$$

To make the evaluation of the integral in Eq. (2) efficient, one can use conjugate models (Xuan & Murphy, 2007) or approximations (Turner et al., 2013; Niekum et al., 2014),

which make the following recursion efficient, too:

$$d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t, m_t) = \sum_{m_{t-1}} \sum_{r_{t-1}} \Big\{ d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_t, m_t)\mathbb{P}(m_t|\boldsymbol{y}_{1:(t-1)}, r_t, m_{t-1}) \mathbb{P}(r_t|r_{t-1}) d\mathbb{P}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}. \quad (3)$$

The recursion in AM is the special case for $|\mathcal{M}| = 1$. For $|\mathcal{M}| > 1$, $\mathbb{P}(m_t|m_{t-1}, r_t, \boldsymbol{y}_{1:(t-1)})$ arises as a new term, which for $\mathbb{1}_a$ as the indicator function of a is given by

$$\begin{cases} \mathbb{1}_{m_{t-1}}(m_t)\mathbb{P}(m_{t-1}|\boldsymbol{y}_{1:(t-1)}, r_t) & \text{if } r_t > 0 \\ q(m_t) & \text{if } r_t = 0. \end{cases} \quad (4)$$

Next, define the *growth-* and *changepoint probabilities* as

$$d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = r_{t-1} + 1, m_t) = d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_t, m_t) d\mathbb{P}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \times (1 - H(r_t))\mathbb{P}(m_{t-1}|\boldsymbol{y}_{1:(t-1)}, r_t), \quad (5a)$$

$$d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = 0, m_t) = d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{y}_{1:(t-1)}, r_t, m_t) q(m_t) \times \sum_{m_{t-1}} \sum_{r_{t-1}} \Big\{ H(r_{t-1} + 1) d\mathbb{P}(\boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}. \quad (5b)$$

The evidence can then be calculated via Eq. (6a), which in turn allows calculating the joint model-and-run-length distribution (6b), the model posterior (6c), as well as the model-specific (6d) and global (6e) run-length distributions:

$$d\mathbb{P}(\boldsymbol{y}_{1:t}) = \textstyle\sum_{m_t} \sum_{r_t} d\mathbb{P}(\boldsymbol{y}_{1:t}, m_t, r_t) \quad (6a)$$

$$\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t}) = d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t, m_t)/d\mathbb{P}(\boldsymbol{y}_{1:t}) \quad (6b)$$

$$\mathbb{P}(m_t|\boldsymbol{y}_{1:t}) = \textstyle\sum_{r_t} \mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t}) \quad (6c)$$

$$\mathbb{P}(r_t|m_t, \boldsymbol{y}_{1:t}) = d\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t})/d\mathbb{P}(m_t|\boldsymbol{y}_{1:t}) \quad (6d)$$

$$\mathbb{P}(r_t|\boldsymbol{y}_{1:t}) = \textstyle\sum_{m_t} \mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t}) \quad (6e)$$

$$\mathbb{P}(m_{t-1}|\boldsymbol{y}_{1:(t-1)}, r_t) = \frac{\mathbb{P}(m_{t-1}, r_{t-1}|\boldsymbol{y}_{1:(t-1)})}{\mathbb{P}(r_{t-1}|\boldsymbol{y}_{1:(t-1)})}. \quad (6f)$$

Eq. (6f) is the conditional model posterior from Eq. (4). Eq. (6e) is arrived at directly in FL and used for on-line MAP segmentation. By framing our derivations in the run-length framework of AM, we additionally obtain (4)–(6d), thus enabling on-line prediction and model selection at the same computational cost.

2.3. On-line algorithm outputs

Prediction: Recursive h-step-ahead forecasting uses (6b):

$$d\mathbb{P}(\boldsymbol{Y}_{t+h}|\boldsymbol{y}_{1:t}) = \sum_{r_t} \sum_{m_t} \Big\{ d\mathbb{P}(\boldsymbol{Y}_{t+h}|\boldsymbol{y}_{1:t}, \widehat{\boldsymbol{y}}_t^h, r_t, m_t)\mathbb{P}(r_t, m_t|\boldsymbol{y}_{1:t}) \Big\}, \quad (7)$$

where $\widehat{\boldsymbol{y}}_t^h = \emptyset$ if $h = 1$ and $\widehat{\boldsymbol{y}}_t^h = \widehat{\boldsymbol{y}}_{(t+1):(t+h-1)}$ otherwise, with $\widehat{\boldsymbol{y}}_{t+h} = \mathbb{E}(\boldsymbol{Y}_{t+h}|\boldsymbol{y}_{1:t}, \widehat{\boldsymbol{y}}_t^h)$ the recursive forecast.

Tracking the model posterior/Bayes Factors: One of the novel capabilites of the algorithm is on-line monitoring of the model posterior via Eq. (6c). This is attractive when structural changes in the data happen slowly and are not captured well by CPs. In this case, $\mathbb{P}(m_t|\boldsymbol{y}_{1:t})$ can be used to identify periods of change, see Fig. 6. For pairwise comparisons, Bayes Factors can be monitored, too:

$$\text{BF}(m_1, m_2)_t = \frac{\mathbb{P}(m_t = m_1|\boldsymbol{y}_{1:t}) \cdot q(m_2)}{\mathbb{P}(m_t = m_2|\boldsymbol{y}_{1:t}) \cdot q(m_1)}. \quad (8)$$

Maximum A Posteriori (MAP) segmentation: For MAP_t the density of the MAP-estimate at t and $\text{MAP}_0 = 1$, FL's recursive MAP estimator is given by

$$\text{MAP}_t = \max_{r,m} \Big\{ d\mathbb{P}(\boldsymbol{y}_{1:t}, r_t = r, m_t = m)\text{MAP}_{t-r-1} \Big\}. \quad (9)$$

For r_t^*, m_t^* maximizers for time t, the MAP segmentation is $S_t = S_{t-r_t^*-1} \cup \{(t - r_t^*, m_t^*)\}$, $S_0 = \emptyset$, where $(t', m_{t'}) \in S_t$ means a CP at $t' \leq t$, with $m_{t'} \in \mathcal{M}$ the model for $\boldsymbol{y}_{t':t}$.

3. Building a spatio-temporal model universe

The last section derived BOCPDMS for arbitrary data streams $\{\boldsymbol{Y}_t\}$. Next, we propose models for $\mathcal{M}$ if $\{\boldsymbol{Y}_t\}$ can be mapped into a space $\mathbb{S}$. Let $\mathcal{S}$ with $|\mathcal{S}| = S$ be a set of spatial locations in $\mathbb{S}$ with measurements $\boldsymbol{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,S})^T$ recorded at times $t = 1, 2, \ldots$

3.1. Bayesian VAR (BVAR)

Inference on $\{\boldsymbol{Y}_t\}$ can be drawn using conjugate Bayesian Vector Autoregressions (BVAR) with lag length L and E additional variables $\boldsymbol{Z}_t$ as elements of model universe $\mathcal{M}$:

$$\sigma^2 \sim \text{InverseGamma}(a, b) \quad (10a)$$

$$\boldsymbol{\varepsilon}_t|\sigma^2 \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \cdot \boldsymbol{\Omega}) \quad (10b)$$

$$\boldsymbol{c}|\sigma^2 \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \cdot \boldsymbol{V}_c) \quad (10c)$$

$$\boldsymbol{Y}_t = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{Z}_t + \textstyle\sum_{l=1}^{L} \boldsymbol{A}_l \boldsymbol{Y}_{t-l} + \boldsymbol{\varepsilon}_t. \quad (10d)$$

Here, $\boldsymbol{A}_l, \boldsymbol{B}$ are $S \times S$, $S \times E$ matrices, $\boldsymbol{c} = (\boldsymbol{\alpha}, \text{vec}(\boldsymbol{B}), \text{vec}(\boldsymbol{A}_1), \text{vec}(\boldsymbol{A}_2), \ldots \text{vec}(\boldsymbol{A}_L))^T$ is a vector of $S \cdot (LS + 1 + E)$ model parameters. Scalars $a, b > 0$, matrix $\boldsymbol{V}_c$, and diagonal matrix $\boldsymbol{\Omega}$ are hyperparameters.

3.2. Approximating processes using VAR

Modelling $\{\boldsymbol{Y}_t\}$ as VAR is attractive, as many complex non-linear processes have VAR representations, including HMMs, time-stationary GPs as well as multivariate GARCH and fractionally integrated VARMA processes (Inoue & Kasahara, 2006; Inoue et al., 2018). Performance guarantees for VAR approximations to such processes are derived using Baxter's Inequalitiy with multivariate versions of results in Hannan & Kavalieris (1986).

Theorem 1. *Let $\{\boldsymbol{Y}_t\}$ be a time-stationary spatio-temporal process with spectral density satisfying regularity condition* ***(A)*** *of Meyer & Kreiss (2015), $||\cdot||$ a matrix norm, $\mathbb{E}(\boldsymbol{Y}_t) = 0$, $\mathbb{E}(\boldsymbol{Y}_t\boldsymbol{Y}_t^T) < \infty$, $\sum_{h=-\infty}^{\infty}(1+|h|)^3||\mathbb{E}[\boldsymbol{Y}_t\boldsymbol{Y}'_{t+h}]|| < \infty$. Then (1)–(3) hold.*

(1) $\boldsymbol{Y}_t = \sum_{i=1}^{\infty}\boldsymbol{A}_i\boldsymbol{Y}_{t-i} + \boldsymbol{\varepsilon}_t$ for matrices $\{\boldsymbol{A}_l\}_{l\in\mathbb{N}}$ and $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$, $\mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t) = \boldsymbol{D}$, $\boldsymbol{D}$ diagonal.

(2) For $\boldsymbol{Y}_t = \sum_{l=1}^{L}\boldsymbol{A}_l^L\boldsymbol{Y}_{t-l} + e_t$ with $\{\boldsymbol{A}_l^L\}_{l=1}^L$ the best linear projection coefficients, $\exists L_0 : \forall L > L_0$, $\sum_{l=1}^{L}(1+|l|)^3||\boldsymbol{A}_l^L - \boldsymbol{A}_l|| \leq C\cdot\sum_{l=L+1}^{\infty}(1+|l|)^3||\boldsymbol{A}_l||$ with C constant.

(3) Using T observations with $L = \mathcal{O}([T/\ln(T)]^{1/6})$ to estimate $\boldsymbol{A}_l^L$ as the MAP $\widehat{\boldsymbol{A}}_l^L$ of (10a)–(10d)*, it holds that $L(T)^2\sum_{l=1}^{L(T)}||\widehat{\boldsymbol{A}}_l^{L(T)} - \boldsymbol{A}_l^{L(T)}|| \xrightarrow{P} 0$ as $T\to\infty$.*

Proof. Part (1) is shown in Inoue et al. (2018), part (2) in Lemma 3.1 of Meyer & Kreiss (2015). Part (3) follows by their Remark 3.3 if we can prove that the MAP estimator $\hat{\boldsymbol{c}}(L(T))$ of $\boldsymbol{c}$ equals its Yule-Walker estimator (YWE) as $T\to\infty$. Let $\boldsymbol{B} = \boldsymbol{0}$, $\boldsymbol{\alpha} = \boldsymbol{0}$ and note that YWE equals OLS as $T\to\infty$. With $\boldsymbol{X}_{1:T}$ the regressor matrix of $\boldsymbol{Y}_{t-L(T):t}$, $\hat{\boldsymbol{c}}(L(T)) = (\boldsymbol{X}'_{1:T}\boldsymbol{X}_{1:T} + \boldsymbol{V}_c^{-1})^{-1}(\boldsymbol{X}'_{1:T}\boldsymbol{Y}_{1:T})$. Then, part (3) holds as OLS $\xrightarrow{P}$ $\mathbb{E}(\boldsymbol{X}'_{1:T}\boldsymbol{X}_{1:T})^{-1}\mathbb{E}(\boldsymbol{X}'_{1:T}\boldsymbol{Y}_{1:T})$ and

$$\begin{aligned}\hat{\boldsymbol{c}}(L(T)) &= (\boldsymbol{X}'_{1:T}\boldsymbol{X}_{1:T} + \boldsymbol{V}_c^{-1})^{-1}(\boldsymbol{X}'_{1:T}\boldsymbol{Y}_{1:T})\\ &= (\frac{1}{T}\boldsymbol{X}'_{1:T}\boldsymbol{X}_{1:T} + \frac{1}{T}\boldsymbol{V}_c^{-1})^{-1}\frac{1}{T}(\boldsymbol{X}'_{1:T}\boldsymbol{Y}_{1:T})\\ &\xrightarrow{P} \mathbb{E}(\boldsymbol{X}'_{1:T}\boldsymbol{X}_{1:T})^{-1}\mathbb{E}(\boldsymbol{X}'_{1:T}\boldsymbol{Y}_{1:T}).\end{aligned}$$ □

In Thm. 1, assuming $\mathbb{E}(\boldsymbol{Y}_t) = \boldsymbol{0}$ is without loss of generality: If $\mathbb{E}(\boldsymbol{Y}_t) = \boldsymbol{\alpha}+\boldsymbol{B}\boldsymbol{Z}_t$, define $\boldsymbol{Y}_t^* = \boldsymbol{Y}_t - (\boldsymbol{\alpha}+\boldsymbol{B}\boldsymbol{Z}_t)$ and apply the theorem to $\{\boldsymbol{Y}_t^*\}$. Moreover, the results do *not* require stationarity in space. Lastly, part (3) suggests a principled way of picking lag lengths $\mathcal{L}$ for BVAR models: If between T_1 and T_2 observations are expected between CPs, $\mathcal{L} = \{L\in\mathbb{N} : L(T_1)\leq L\leq L(T_2)\}$. In the experiments of section 6, we employ this strategy using $T_1 = 1, T_2 = T$.

3.3. Modeling spatial dependence

While Thm. 1 motivates approximating spatio-temporal processes between CPs with (10a)–(10d), the matrices $\{\boldsymbol{A}_l^L\}_{l=1}^L$ have $S(LS+1+E)$ parameters. This increases model complexity and ignores spatial information. We remedy both issues through neighbourhood systems on $\mathcal{S}$.

Definition 1 (Neighbourhood system). *Let $\mathcal{S}$ be all locations. Define for all $s\in\mathcal{S}$ the sets $N_i(s)\subseteq\mathcal{S}$ with $0\leq i\leq n$ as the i-th* neighbourhood of s*, so that $N_i(s)\cap N_j(s) = \emptyset$, $s'\in N_i(s) \iff s\in N_i(s')$ and $N_0(s) = \{s\}$. Finally, define the* neighbourhood system *as $N(\mathcal{S}) = \{\{N_i(s)\}_{i=1}^n : s\in\mathcal{S}, 0\leq i\leq n\}$.*

In the remainder of the paper, smaller indices i imply that the neighbourhoods $N_i(s)$ are closer to s. For a BVAR model of lag length L, the decay of spatial dependence is encapsulated through $p : \{1,\dots,L\}\to\{0,\dots,n\}$. In particular, only $s'\in N_i(s)$ with $i\leq p(l)$ are modeled as affecting s after l time periods.

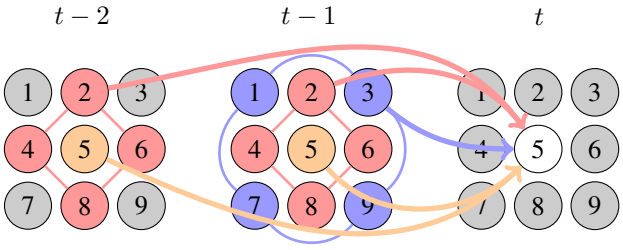

Figure 2. SSBVAR modeling: Suppose that on a regular grid of size 9, $Y_{t,5}$ depends on the past two realizations of itself and its 4- neighbourhood, and the last realization of its 8-neighbourhood. This is an SSBVAR on $\mathcal{S} = \{1,\dots,9\}$ with $L = 2$, $N_0(5) = \{5\}$, $N_1(5) = \{2,4,6,8\}$, $N_2(5) = \{1,3,7,9\}$ and function p with $p(1) = 2, p(2) = 1$.

3.4. Spatializing BVAR

In principle, given $N(\mathcal{S})$, sparsification of the BVAR model (10a)–(10d) is possible in two ways: As restriction on the *contemporaneous* dependence via the covariance matrix of the error term $\boldsymbol{\varepsilon}_t$, or as restriction on the *conditional* dependence via the coefficient matrices $\{\boldsymbol{A}_l\}_{l=1}^L$. We choose the latter for three reasons: Firstly, linear effects have more interesting interpretations than error covariances. Secondly, using $\{\boldsymbol{A}_l\}_{l=1}^L$ to encode spatial dependency allows us to work with arbitrary neighbourhoods. In contrast, modelling dependent errors under conjugacy limits dependencies to decomposable graphs (Xuan & Murphy, 2007). Since not even a regular grid is decomposable, this is problematic for spatial data. Thirdly, modelling errors as contemporaneous is attractive for low-frequency data where the resolution of temporal effects is coarse, but the situation reverses for high-frequency data. Since the algorithm runs on-line, we expect $\{\boldsymbol{Y}_t\}$ to be observed with high frequency.

Definition 2 (Spatially structured BVAR (SSBVAR)). *For process $\{\boldsymbol{Y}_t\}$ on $\mathcal{S}$ and $(L, N(\mathcal{S}), p(\cdot))$, define the matrices $\{\widetilde{\boldsymbol{A}}_l\}_{l=1}^L$ by imposing that $[\widetilde{\boldsymbol{A}}_l]_{(s,s')} = 0 \iff s'\notin N_i(s)$ for any $i\leq p(l)$. Let $\widetilde{\boldsymbol{A}}_l^{\neq 0}$ be the vector of non-zero entries in $\widetilde{\boldsymbol{A}}_l$ and $\widetilde{\boldsymbol{c}} = (\boldsymbol{\alpha}, vec(\boldsymbol{B}), \widetilde{\boldsymbol{A}}_1^{\neq 0}, \widetilde{\boldsymbol{A}}_2^{\neq 0}, \dots \widetilde{\boldsymbol{A}}_L^{\neq 0})^T$. The SSBVAR model on $\{\boldsymbol{Y}_t\}$ induced by $(L, N(\mathcal{S}), p(\cdot))$ is obtained by combining* (10a)–(10b) *with*

$$\widetilde{\boldsymbol{c}}|\sigma^2 \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\cdot\boldsymbol{V}_{\widetilde{c}}) \qquad (10e)$$

$$\boldsymbol{Y}_t = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{Z}_t + \sum_{l=1}^{L}\widetilde{\boldsymbol{A}}_l\boldsymbol{Y}_{t-l} + \boldsymbol{\varepsilon}_t. \qquad (10f)$$

Fig. 2 illustrates this idea. Further sparsification is possible by modelling neighbourhoods jointly, i.e. $[\widetilde{\boldsymbol{A}}_l]_{(s,s')} =$

$a_i(s), \forall s' \in N_i(s)$, reducing the number of parameters to $S \cdot \sum_{l=1}^{L} p(l)$. If one imposes $a_i(s) = a_i(s') = \cdots = a_i$, this number drops to $\sum_{l=1}^{L} p(l)$.

3.5. Building SSBVAR: choosing $L, N(\mathcal{S}), p(\cdot)$

For the choice of lag lengths L, part *(3)* of Thm. 1 suggests $L \in \{L' \in \mathbb{N} : L(T_1) \leq L' \leq L(T_2)\}$ if one expects T_1 to T_2 observations between CPs. For any data stream $\{\boldsymbol{Y}_t\}$ on a space $\mathbb{S}$, there are different ways of constructing neighbourhood structures $N(\mathcal{S})$. For example, when analysing pollutants in London's air in section 6, $N(\mathcal{S})$ could be constructed from Euclidean or Road distances. By filling $\mathcal{M}$ with SSBVARs constructed using competing versions of $N(\mathcal{S})$, BOCPDMS provides a way of dealing with such uncertainty about spatial relations. In fact, it can dynamically discern changing spatial relationships on $\mathbb{S}$. Lastly, $p(\cdot)$ should be decreasing, reflecting that measurements affect each other less when further apart.

4. Hyperparameter optimization

Hyperparameter inference on $\boldsymbol{\theta}_m^0$ can be adressd either by introducing an additional hierarchical layer (Wilson et al., 2010) or using type-II ML. The latter is obtained by maximizing the model-specific evidence

$$\log d\mathbb{P}(\boldsymbol{y}_{1:T}|\boldsymbol{\theta}_m^0) = \sum_{t=1}^{T} \log d\mathbb{P}(\boldsymbol{y}_t|\boldsymbol{\theta}_m^0, \boldsymbol{y}_{1:(t-1)}). \quad (11)$$

Computation of the righthand side requires evaluating the gradients $\nabla\theta_m^0 dP(\boldsymbol{y}_{1:t}, r_t|\boldsymbol{\theta}_m^0)$. These are efficiently obtained using recursions similar to those in section 2. Turner et al. (2009) use the first T' observations as a test set, and employ conjugate gradient descent by running BOCPD K times on $\boldsymbol{y}_{1:T'}$ to find $\widehat{\boldsymbol{\theta}}_m^0 = \arg\min_{\boldsymbol{\theta}_m^0} \left\{d\mathbb{P}(\boldsymbol{y}_{1:T'}|\boldsymbol{\theta}_m^0)\right\}$. In contrast, Caron et al. (2012) propose on-line gradient descent updates via

$$\boldsymbol{\theta}_{m,t+1}^0 = \boldsymbol{\theta}_{m,t}^0 + \alpha_t \nabla \boldsymbol{\theta}_{m,t+1}^0 \log d\mathbb{P}(\boldsymbol{y}_{t+1}|\boldsymbol{\theta}_{m_{1:t}}^0, \boldsymbol{y}_{1:t}).$$

The latter is preferable for two reasons: Firstly, inference and type-II ML are executed simultaneously (rather than sequentially) and thus enable cold-starts of BOCPDMS. Secondly, neither the on-line nature nor the computational complexity of BOCPDMS is affected.

5. Computation & Complexity

In spite of tracking $|\mathcal{M}|$ models, BOCPDMS has linear time complexity. Step 1 in the pseudocode is the computationally most demanding one, but the loop over $\mathcal{M}$ can be parallelized: With N threads, it executes in $\mathcal{O}\left(\lceil|\mathcal{M}|/N\rceil \cdot \max_{M\in\mathcal{M}} \text{CompTime}(M)\right)$. Step 2 takes $\mathcal{O}(|R(t)||\mathcal{M}|)$, for $R(t)$ the set of run-lengths at time t.

5.1. Pruning the run-length distribution

In a naive implementation, all run-lengths are retained and $R(t) = \{1, 2, \ldots, t\}$. This implies execution time of order $\mathcal{O}(t)$ for processing $\boldsymbol{y}_t$. However, this can be made time-constant by pruning the run-length distribution: If one discards run-lengths whose probability is $\leq 1/R_{\max}$ or only keeps the $R_{\max}$ most probable ones, $|R(t)| \leq R_{\max}$(Adams & MacKay, 2007). A third way is *Stratified Rejection Control (SRC)* (Fearnhead & Liu, 2007), which Caron et al. (2012) and the current paper found to perform as well as the other approaches. In our setting, we prune based on the model-specific run-length distribution $\mathbb{P}(r_t|m_t, \boldsymbol{y}_{1:t})$ rather than on $\mathbb{P}(r_t|\boldsymbol{y}_{1:t})$.

5.2. BVAR updates

The bottleneck when updating an (SS)BVAR model in $\mathcal{M}$ is step I.B.1 in the pseudocode of BOCPDMS, when updating the MAP estimate $\boldsymbol{c}(r,t) = \boldsymbol{F}(r,t)\boldsymbol{W}(r,t)$ of the coefficient vector , where $\boldsymbol{F}(r,t) = (\boldsymbol{X}'_{(t-r):t}\boldsymbol{X}_{(t-r):t} + \boldsymbol{V}_{\overline{c}})^{-1}$ and $\boldsymbol{W}(r,t) = \boldsymbol{X}'_{(t-r):t}\boldsymbol{Y}_{(t-r):t}$ for all $r \in R(t)$. Since $\boldsymbol{W}(r,t) = \boldsymbol{W}(r-1,t-1) + \boldsymbol{X}_t'\boldsymbol{Y}_t$, updates are $\mathcal{O}(kS)$. $\boldsymbol{F}(r-1,t-1)$ can be updated to $\boldsymbol{F}(r,t)$ using rank-k updates to its QR-decomposition in $\mathcal{O}(k^3)$ or using Woodbury's formula, in $\mathcal{O}(S^3)$, implying an overall complexity of $\mathcal{O}(|R(t)|\min\{k^3, S^3\})$ at time t.

5.3. Comparison with GP-based approaches

Define $k_{\max}$ as the largest number of regressors of any (SS)BVAR model in $\mathcal{M}$. From the previous paragraphs, it follows that if all models in $\mathcal{M}$ are BVARs, the overhead $C = \lceil N/|\mathcal{M}|\rceil \cdot \min\{k_{\max}^3, S^3\}$ is time-constant. Thus, BOCPDMS runs in $\mathcal{O}(TR_{\max})$ on T observations. In contrast, the models of Saatçi et al. (2010) run in $\mathcal{O}(TR_{\max}^3)$. Even taking into account the constant C, this implies that for $R_{\max} = 100$, it can process a more than 20-dimensional data stream faster than the GP-model a univariate one.

6. Experimental results

We evaluate the performance of the algorithm on real data in two parts. First, we compare it to benchmark performances of GP-based models on real world data reported by Saatçi et al. (2010). This is done to demonstrate the practical implications of Thm. 1, namely that (V)ARs are excellent approximations for a large variety of data streams. Next, we investigate the multivariate setting to showcase BOCPDMS' novelty. All computations can be reproduced with the code available from XXXXXXX and use a uniform model prior q, a constant Hazard function H and use gradient descent for hyperparameter optimization as in Section 4. The lag lengths of models in $\mathcal{M}$ will be chosen based on Thm. 1, part (3) for (SS)BVARs and the rates of Hannan &

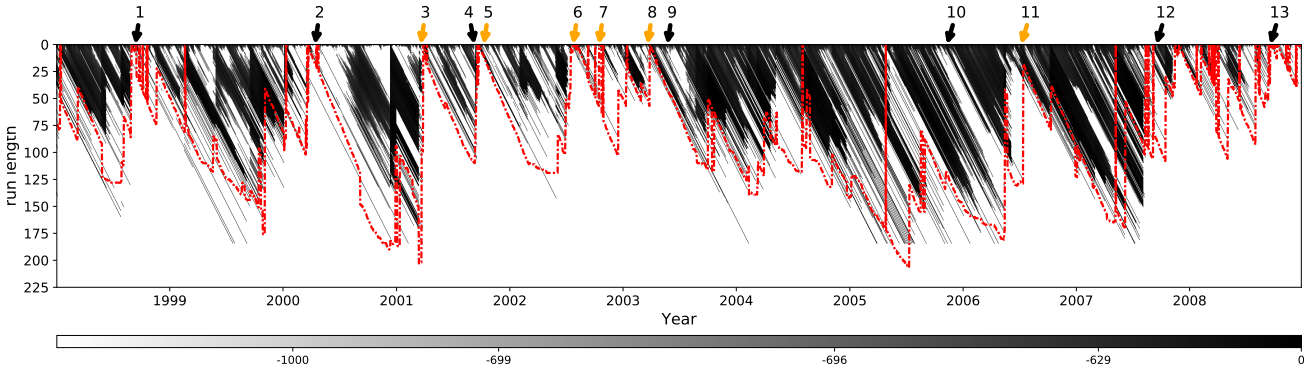

Figure 3. Results for 30 *Portfolio data set, displayed from 01/01/1998–31/12/2008*: Log run-length distribution (grayscale) and its maximum (red). Changepoints (CPs) found by Saatçi et al. (2010) are marked in **black**, additional CPs found by BOCPDMS in orange. Labels correspond to: (1) Asia Crisis, (2) DotCom bubble bursting, (3) OPEC cuts output by 4%, (4) 9/11, (5) Afghanistan war, (6) 2002 stock market crash, (7) Bombing attack in Bali, (8) Iraq war, (9) Major tax cuts under Bush, (10) US election, (11) Iran announces successful enrichment of Uranium, (12) Northern Rock bank run, (13) Lehman Brothers collapse.

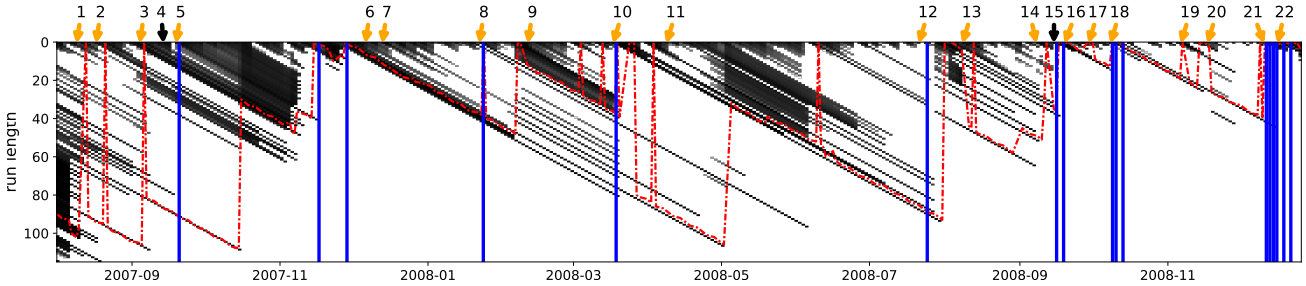

Figure 4. Financial crisis 01/08/2007–31/12/2008: Colours as in Fig 3, with MAP-segmentation (blue). Event labels: (1) BNP Paribas funds frozen, (2) Fed cuts lending rate, (3) IKB 1bn$ losses, (4) Northern Rock bank run, (5) Fed cuts interest rate, (6) Bush rescue plan for $>10^6$ homeowners, (7) Fed, ECB, BoE loans for banks, (8) Fed cuts funds rate, (9) G7 estimate: 400bn$ losses worldwide, (10) JP Morgan buys Bear Stearns, (11) IMF estimate: >1trn$ losses worldwide, (12) HBOS' rights issue fails, (13) ECB provides €200bn for liquidity, (14) Fannie Mae & Freddie Mac bailout, (15) Lehman collapse, (16) Russia: 500bn Roubles crisis package, (17) Fortis bailout, (18) UK: £500bn bank rescue package, (19) BoE, ECB cut interest rate, (20) G20 promise fiscal stimuli, (21) Madoff's Ponzi scheme revealed, South Korean CB sets interest rate at record low (22) Fed, Japanese central bank cut interest rates. Dates from Guillén (2009).

Kavalieris (1986) for Bayesian Autoregressions (BARs).

6.1. Comparison with GP-based approaches

As in Saatçi et al. (2010), ARGPCP will refer to the non-linear GP-based AR model, GPTSCP to the time-deterministic model, and NSGP to the non-stationary GP allowing hyper-parameters to change at every CP. Saatçi et al. (2010) compute the mean squared error (MSE) as well as the negative log predictive likelihood (NLL) of the one-step-ahead predictions for three data sets: The water height of the Nile between $622-1284$ AD, the snowfall in Whistler (Canada) over a 37 year period and the 3-dimensional time series (x-, y-coordinate and headangle) of a honey bee during a waggle dance sequence. In Turner (2012), all of the models except NSGP were also compared on daily returns for 30 industry portfolios from $1975-2008$. In Table 1, BOCPDMS is compared to these benchmarks for $\mathcal{M}$ consisting of BAR and (SS)BVAR models.

6.1.1. Designing $\mathcal{M}$

The Nile and the snowfall data are univariate, meaning $\mathcal{M}$ contains only BARs. For the 3-dimensional bee data, $\mathcal{M}$ contains unrestricted BVARs and BARs. Lastly, different neighbourhood systems are built for the 30 Portfolio data set. Two by computing the spaces of the pairwise contemporaneous correlations and autocorrelations *prior* to 1975 and using distances in these spaces, the third by using the *Standard Industrial Classification (SIC)* system, with $p(\cdot)$ decreasing linearly.

6.1.2. Findings

Predictive performance and fit: In terms of MSE, BOCPDMS clearly outperforms all GP-models on multivariate data. Even on univariate data, the only exception to this is the snowfall data, where NSGP does better. However, NSGP requires grid search or Hamiltonian Monte Carlo

Table 1. Predictive MSE and NLL of BOCPDMS in comparison to GP-based techniques, with 95% error bars. GP results are as reported in Saatçi et al. (2010) and Turner (2012). NLLs marked * are centered in Turner (2012) so that the NLL of the best-performing GP-method amongst them is 0.0.

	NILE		SNOWFALL	
METHOD	MSE	NLL	MSE	NLL
ARGPCP	0.553	1.15	0.750	−0.604
	(0.0962)	(0.0555)	(0.0315)	(0.0385)
GPTSCP	0.583	1.19	0.689	1.17
	(0.0989)	(0.0548)	(0.0294)	(0.0183)
NSGP	0.585	1.15	**0.618**	**−1.98**
	(0.0988)	(0.0655)	(0.0242)	(0.0561)
BVAR	**0.550**	**1.13**	0.681	0.923
	(0.0948)	(0.0684)	(0.0245)	(0.0231)
	BEE DANCE		30 PORTFOLIOS	
METHOD	MSE	NLL	MSE	NLL
ARGPCP	2.62	4.07	29.95	0.17*
	(0.195)	(0.150)	(0.50)	(0.22)
GPTSCP	3.13	4.54	30.17	0.00*
	(0.241)	(0.188)	(0.51)	(0.22)
NSGP	3.17	4.19	–	–
	(0.230)	(0.212)	–	–
BVAR	**1.74**	**3.57**	**25.93**	48.32
	(0.222)	(0.166)	(0.906)	(0.964)

for hyperparameter optimization at each observation (Saatçi et al., 2010). Overall, there are three main reasons why BOCPDMS performs better: Firstly, being able to change lag lengths between CPs seems more important to predictive performance than being able to model non-linear dynamics. Secondly, unlike the GP-models, we allow the time series to communicate via $\{\boldsymbol{A}_l^L\}$. Thirdly, the hyperparameters of the GP have a strong influence on inference. In particular, the noise variance σ is treated as a hyperparameter and optimized via type-II ML. Except for the NSGP, this is only done during a training period. Thus, the GP-models cannot adapt to the observations after training, leading to overconfident predictive distributions that are too narrow (as also noted in Turner, 2012, p. 172). This in turn leads them to be more sensitive to outliers, and to mislabel them as CPs. In contrast, (10a)–(10d) models σ as part of the inferential Bayesian hierarchy, and hyperparameter optimization is instead applied at one level higher. Consequently, our predictive distributions are wider, and the algorithm is less confident about the next observations, making it more robust to outliers. This is also responsible for the overall smaller standard errors of the GP-models in Table 1, since the GPs interpret outliers as CPs and immediately adapt to short-term highs or lows.

CP Detection: A good demonstration of this finding is the Nile data set, where the MAP segmentation finds a single CP, corresponding to the installation of the nilometer around 715 CE, see Fig 5. In contrast, Saatçi et al. (2010) report 18 additional CPs corresponding to outliers. The same phenomenon is also reflected in the run-length distribution (RLD): While the probabilty mass in Figs. 3, 4 and 5 are spread across the retained run-lengths, the RLD reported in Saatçi et al. (2010) is more concentrated and even degenerate for the 30 Portfolio data set. On the other hand, such enhanced sensitivity to change can be advantageous. For instance, in the bee waggle dance, the GP-based techniques are better at identifying the true CPs. The reason is twofold: Firstly, the variance for the bee waggle data is homogeneous across time, so treating it as fixed helps inference. Secondly, the CPs in this data set are subtle, so having narrower predictive distributions is of great help in detecting them. However, it adversely affects performance when changes in the error variance are essential, as for financial data: In particular, BOCPDMS finds the ground truths labelled in Saatçi et al. (2010), and discovers even more, see Fig. 3. This is especially apparent in times of market turmoil where changes in the variance of returns are significant. We show this using the example of the subprime mortgage financial crisis: While the RLD of Saatçi et al. (2010) identified only 2 CPs with ground truth and a third unlabelled one during the height of the crisis, BOCPDMS detects a large number of CPs corresponding to ground truths, see Fig. 4.

Lastly, we note that segmentations obtained off-line for both the bee waggle dance and the 30 Portfolios are reported in Xuan & Murphy (2007). Compared to the on-line segmentations produced by BOCPDMS, these are closer to the truth for the bee waggle data, but not for the 30 Portfolio data set.

Model selection: In most of the experiments where abrupt changes model the non-stationarity well, the model posterior is fairly concentrated and periods of model uncertainty are

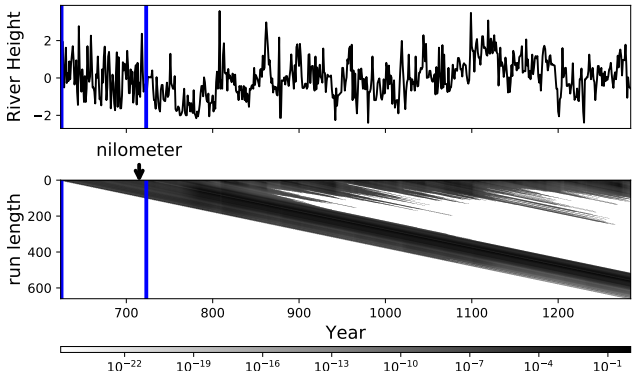

Figure 5. Results for Nile data: **Panel 1:** the data undergoes a structural shift around 715. **Panel 2:** Both run-length distribution (grayscale) and MAP segmentation (blue) detect the change.

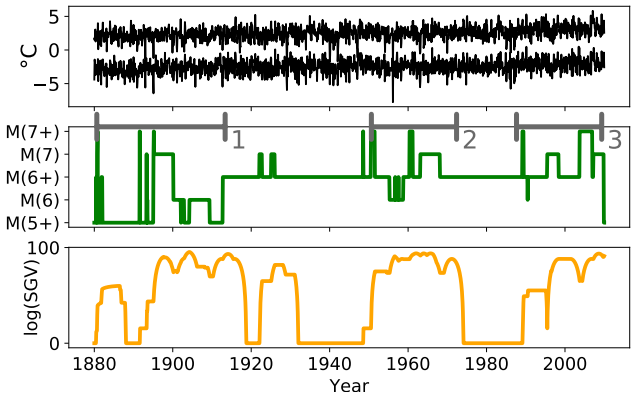

Figure 6. Results for European Temperatures: **Panel 1:** normalized temperature for Prague and Jena **Panel 2:** Model Posterior maximum, $\widehat{m}_t = \arg\max_{\mathcal{M}}\{\mathbb{P}(m_t|y_{1:t})\}$, model complexity decreasing top to bottom. $M(l), M(l+)$ are SSBVAR with l lags. Spatial dependence in $M(l+)$ is slower decaying. Periods of model uncertainty are (1) 2nd Industrial Revolution $1870 - 1914$, (2) Post WW2 boom $1950 - 1973$, (3) European Climate shift $1987-$present, see Luterbacher et al. (2004). **Panel 3:** To compare model uncertainty across different data and $\mathcal{M}$, the (Log) Standardized Generalized Variance (SGV) of $\widehat{m}_t$ can be used.

short. This is different when changes are slower, see Fig. 6. The implicit model complexity penalization Bayesian model selection performs provides BOCPDMS with an Occam's Razor mechanism: Simple models are typically favoured until evidence for more complex dynamics accumulates. For the bee waggle and the 30 Portfolio data set, BVARs are preferred to BARs. For the 30 Portfolio data, the MAP segmentation only selects SSBVARs with neighbourhoods constructed from contemporaneous correlation and autocorrelations. Neighbourhoods using SIC codes are not selected, reflecting that this classification from 1937 is out of date.

6.2. Performance on spatio-temporal data

European Temperature: Monthly temperature averages $01/01/1880 - 01/01/2010$ for the 21 longest-running stations across Europe are taken from http://www.ecad.eu/. We adjust for seasonality by subtracting monthly averages for each station. Station longitudes and latitudes are available, so $N(\mathcal{S})$ is based on concentric rings around the stations using Euclidean distances. Two different decay functions $p(\cdot), p^+(\cdot)$ are used, with $p^+(\cdot)$ using larger neighbourhoods and slower decaying. Temperature changes are poorly modeled by CPs and more likely to undergo slow transitions. Fig. 6 shows the way in which the model posterior captures such longer periods of change in dynamics. The values on the bottom panel are calculated by considering $\widehat{m}_t = \arg\max_{M\in\mathcal{M}} \mathbb{P}(m_t|\boldsymbol{y}_{1:t})$ as $|\mathcal{M}|$-dimensional multinomial random variable. Its Standardized Generalized Variance (SGV) (Wilks, 1960; SenGupta, 1987) is calculated as $|\mathcal{M}|$-th root of the covariance matrix determinant. We plot the log of the SGV computed using the model posteriors for the last 8 years. This provides an informative summary of the model posterior dispersion.

Air Pollution: Finally, we analyze Nitrogen Oxide (NOX) observed at 29 locations across London $17/08/2002 - 17/08/2003$. The quarterhourly measurements are averaged over 24 hours. Weekly seasonality is accounted for by subtracting week-day averages for each station. $\mathcal{M}$ is populated with SSBVAR models whose neighbourhoods are constructed from both road- and Euclidean distances. As $17/02/2003$ marks the introduction of London's first ever congestion charge, we find structural changes in the dynamics around that date. A model with shorter lag length but identical neighbourhood structure is preferred after the congestion charge. In Fig. 7, Bayes Factors (BFs) capture the shift: Kass & Raftery (1995) classify logs of BFs as very strong evidence if their absolute value exceeds 5.

7. Conclusion

We have extended Bayesian On-line Changepoint Detection (BOCPD) to multiple models by generalizing Fearnhead & Liu (2007) and Adams & MacKay (2007), arriving at BOCPDMS. For inference in multivariate data streams, we propose BVARs with closed form distributions that have strong theoretical guarantees summarized in Thm. 1. We sparsify BVARs based on neighbourhood systems, thus making BOCPDMS especially amenable to spatio-temporal inference. To demonstrate the power of the resulting framework, we apply it to multivariate real world data, outperforming the state of the art. In future work, we would like to add and remove models from $\mathcal{M}$ on-line. This could lower the computational cost for the case where $|\mathcal{M}|$ is significantly larger than the number of threads.

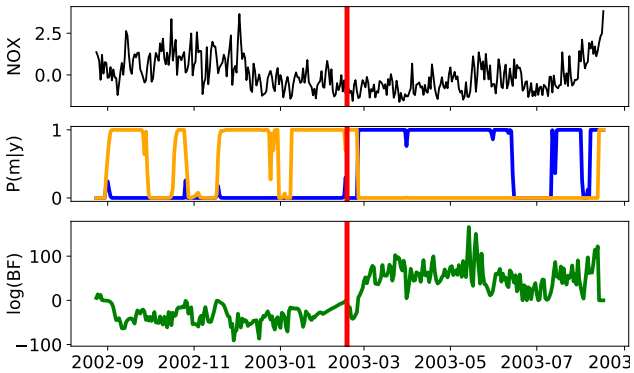

Figure 7. Results for Air Pollution: **Panel 1:** **NOX** levels for Brent, with congestion charge introduction date (red) **Panel 2:** Model posteriors for the two best-fitting models, with Euclidean neighbourhoods. **Panel 3:** Log Bayes Factors of these two models

Acknowledgements

JK is funded by the EPSRC grant EP/L016710/1 . Further, this work was supported by The Alan Turing Institute under EPSRC grant EP/N510129/1 as well as the Lloyds Register Foundation programme on Data Centric Engineering.

References

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

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

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

Fazayeli, Farideh and Banerjee, Arindam. Generalized direct change estimation in ising model structure. In *International Conference on Machine Learning*, pp. 2281–2290, 2016.

Fearnhead, Paul. Exact bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, 53(6):2160–2166, 2005.

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

Garnett, Roman, Osborne, Michael A, and Roberts, Stephen J. Sequential bayesian prediction in the presence of changepoints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 345–352. ACM, 2009.

Guillén, Mauro F. The global economic & financial crisis: A timeline. *The Lauder Institute, University of Pennsylvania*, pp. 1–91, 2009.

Hannan, EJ and Kavalieris, L. Regression, autoregression models. *Journal of Time Series Analysis*, 7(1):27–49, 1986.

Herlands, William, Wilson, Andrew, Nickisch, Hannes, Flaxman, Seth, Neill, Daniel, Van Panhuis, Wilbert, and Xing, Eric. Scalable gaussian processes for characterizing multidimensional change surfaces. In *Artificial Intelligence and Statistics*, pp. 1013–1021, 2016.

Inoue, Akihiko and Kasahara, Yukio. Explicit representation of finite predictor coefficients and its applications. *The Annals of Statistics*, pp. 973–993, 2006.

Inoue, Akihiko, Kasahara, Yukio, Pourahmadi, Mohsen, et al. Baxters inequality for finite predictor coefficients of multivariate long-memory stationary processes. *Bernoulli*, 24(2):1202–1232, 2018.

Kass, Robert E and Raftery, Adrian E. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Khaleghi, Azadeh and Ryabko, Daniil. Asymptotically consistent estimation of the number of change points in highly dependent time series. In *International Conference on Machine Learning*, pp. 539–547, 2014.

Li, Shuang, Xie, Yao, Dai, Hanjun, and Song, Le. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems*, pp. 3366–3374, 2015.

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

Liu, Song, Yamada, Makoto, Collier, Nigel, and Sugiyama, Masashi. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43: 72–83, 2013.

Luterbacher, Jürg, Dietrich, Daniel, Xoplaki, Elena, Grosjean, Martin, and Wanner, Heinz. European seasonal and annual temperature variability, trends, and extremes since 1500. *Science*, 303(5663):1499–1503, 2004.

Meyer, Marco and Kreiss, Jens-Peter. On the vector autoregressive sieve bootstrap. *Journal of Time Series Analysis*, 36(3):377–397, 2015.

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

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

SenGupta, Ashis. Generalizations of barlett's and hartley's tests of homogeneity using overall variability. *Communications in Statistics-Theory and Methods*, 16(4):987–996, 1987.

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

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

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

Wilks, SS. Multidimensional statistical scatter. *Contributions to probability and statistics*, pp. 486–503, 1960.

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

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