UNIVERSITY OF CAMBRIDGE

Gaussian Process Change Point Models

Yunus Saatçi, Ryan Turner and Carl Edward Rasmussen

Computational and Biological Learning

1 Introduction

- Combine Bayesian change point detection with Gaussian Processes to define a nonstationary time series model.
- Central aim is to react to underlying regime changes in an online manner.
- Able to integrate out all latent variables and optimize hyperparameters sequentially.
- Explore three alternative ways of augmenting GP models to handle nonstationarity (GPTS, ARGP-CP and NSGP – see below).
- A Bayesian approach (BOCPD) for online change point detection was introduced in [1].
- BOCPD introduces a latent variable representing the run length at time t and adapts predictions via integrating out the run length.
- BOCPD has two key ingredients:
  - Any model which can construct a predictive density for future observations, in particular, $p(x_t|x_{(t-\tau):(t-1)}, \theta_m)$, i.e., the "underlying predictive model" (UPM).
  - A hazard function $H(r|\theta_h)$ which encodes our prior belief in a change point occuring after observing a run length r.

2 The BOCPD algorithm

Predictions are made robust to change points as follows:

$$p(x_{t+1}|x_{1:t}) = \sum_{r_t} p(x_{t+1}|x_{1:t}, r_t)p(r_t|x_{1:t}) = \sum_{r_t} p(x_{t+1}|x_t^{(r)})p(r_t|x_{1:t}), \quad (1)$$

$$\begin{aligned}\gamma_t := p(r_t, x_{1:t}) &= \sum_{r_{t-1}} p(r_t, r_{t-1}, x_{1:t}) = \sum_{r_{t-1}} p(r_t, x_t|r_{t-1}, x_{1:t-1})p(r_{t-1}, x_{1:t-1}) \\ &= \sum_{r_{t-1}} \underbrace{p(r_t|r_{t-1})}_{\text{hazard}} \underbrace{p(x_t|r_{t-1}, x_t^{(r)})}_{\text{likelihood (UPM)}} \underbrace{p(r_{t-1}, x_{1:t-1})}_{\gamma_{t-1}}. \quad (2)\end{aligned}$$

Defines forward message passing scheme.

3 BOCPD using Gaussian Processes

- GP Time Series (GPTS) UPM gives rise to a Gaussian predictive distribution:

$$p(x_t|x_{(t-\tau):(t-1)}, \lambda) := p(x_t|\mathbf{x}, \lambda) = \mathcal{N}(m_t, v_t), \quad (3)$$

where

$$m_t = \mathbf{k}_\star^\top (K + \sigma_n^2 I)^{-1}\mathbf{x}, \quad (4)$$

$$v_t = k(x_t, x_t) - \mathbf{k}_\star^\top (K + \sigma_n^2 I)^{-1}\mathbf{k}_\star. \quad (5)$$

and λ are the GP hyperparameters.

- The Non-stationary GP (NSGP) UPM handles switches in GP hyperparameters:

$$p(x_t|\mathbf{x}, \phi) = \frac{1}{Z}\int p(x_t|\mathbf{x}, \lambda)p(\mathbf{x}|\lambda)p(\lambda|\phi)\, d\lambda \quad (6)$$

Can approximate the required integral using grid-based methods or Hybrid Monte Carlo.

- The nonstationary autoregressive GP (ARGP-CP) has a UPM implemented by an ARGP predictor:

$$x_t = f(x_{t-p:t-1}) + \epsilon_t, \quad (7)$$

$$f \sim \mathcal{GP}(0, k),\ \epsilon_t \sim \mathcal{N}(0, \sigma^2).$$

- GPTS and NSGP can operate both in continuous and discrete time. ARGP is restricted to discrete time.
- Discrete-time GPTS and NSGP models can be sped up significantly using Toeplitz matrix methods.

4 Hyper-parameter Learning

- BOCPD hyperparameters can be optimized w.r.t. the marginal likelihood (ML) of the entire model.
- Derivatives of the UPM $\frac{\partial}{\partial\theta_m}p(x_t|r_{t-1}, x_t^{(r)}, \theta_m)$, and those of the hazard function, $\frac{\partial}{\partial\theta_h}p(r_t|r_{t-1}, \theta_h)$, can be propagated via a forward message passing scheme that parallels Equations 1 and 2.
- This gives online access to the derivatives of the ML which can be used to adjust such parameters on-the-fly.

5 Results

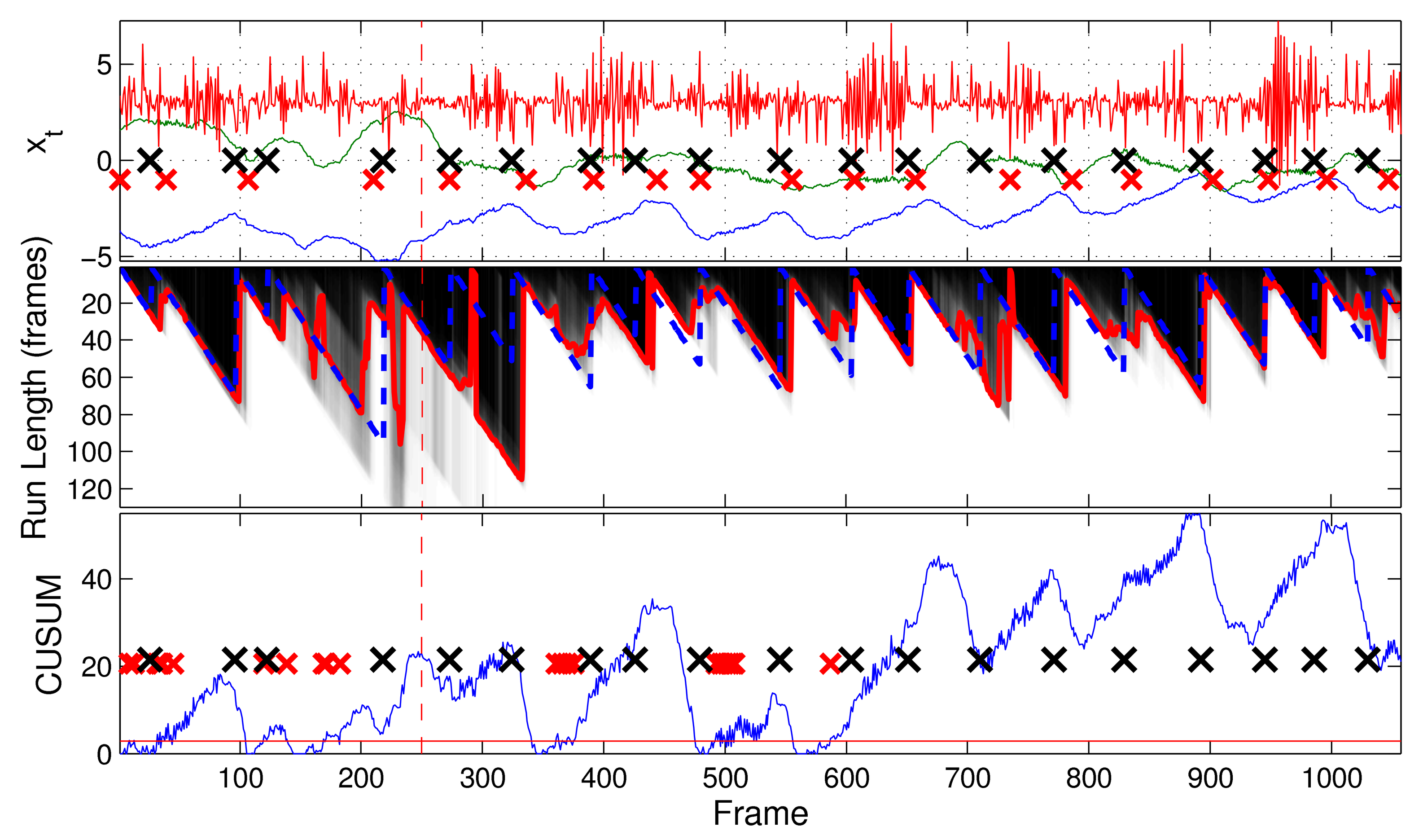

FIGURE 1: **Bee Waggle Dance**: Video sequence one of six. **Top**: The time series are the bee's: x coordinate (blue), y coordinate (green), and angle (red). The vertical dashed red line represents the boundary between train and test sets. **Middle**: True run length: dashed blue. Posterior median in solid red. Large black crosses: the labeled change points. The small red crosses: locations where the probability of a change point > 0.95. **Bottom**: Solid blue line: CUSUM statistic, Horizontal solid red line: the alert threshold set to attain a false positive rate of 5%, Red crosses: Changepoints inferred by CUSUM.

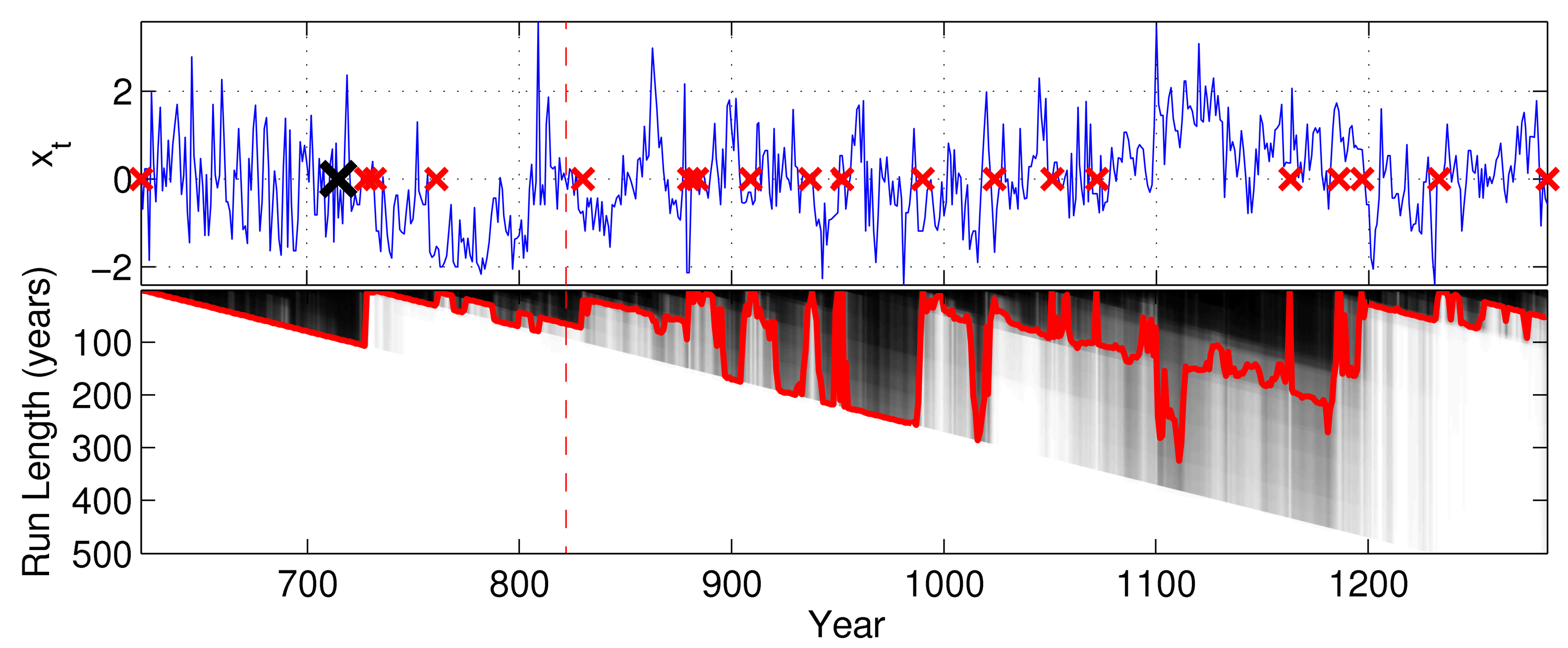

FIGURE 2: **Nile Record:** Between years 622 and 1284. **Top**: The large black cross marks the installation of the nilometer in 715. The small red crosses mark alert locations where the probability of a change point > 0.50. **Bottom**: The run length CDF and its median (solid red).

FIGURE 3: The results are augmented with 95% error bars and the p-value testing the null hypothesis that methods are equivalent to the best performing method, according to NLL, using a one sided t-test.

Method	Negative Log Likelihood	p-value	MSE	p-value
Nile Data (200 Training Points, 462 Test Points)				
GPTS	1.19±0.0548	0.196	0.579±0.0976	0.356
♡ GPTS-CP	1.19±0.0548	0.167	0.583±0.0989	0.335
ARGP	1.18±0.0510	0.202	0.568±0.0940	0.410
♡ ARGP-CP	**1.15 ± 0.0555**	N/A	**0.553 ± 0.0962**	N/A
Kalman	1.17±0.0508	0.361	0.562±0.121	0.453
TIM	1.49±0.0714	<0.001	1.16±0.161	<0.001
♡ NSGP (grid)	1.15±0.0655	0.490	0.585±0.0988	0.321
Bee Waggle Dance Data (250 Training Points, 807 Test Points)				
GPTS	8.02±0.504	<0.001	8.44±0.745	<0.001
♡ GPTS-CP	4.54±0.188	<0.001	3.13±0.241	<0.001
ARGP	4.35±0.167	0.007	2.98±0.224	0.008
♡ ARGP-CP	**4.07 ± 0.150**	N/A	**2.62 ± 0.195**	N/A
Kalman	4.39±0.176	0.002	2.93±0.215	0.016
TIM	4.54±0.177	<0.001	3.25±0.237	<0.001
♡ NSGP (HMC)	4.19±0.212	<0.001	3.17±0.230	<0.001

References

[1] R. P. Adams and D. J. C. MacKay, "Bayesian online changepoint detection," Tecnical Report, University of Cambridge, Cambridge, UK, 2007. arXiv:0710.3742v1 [stat.ML].