Sequential Gaussian Process Prediction in the Presence of Changepoints or Faults

Roman Garnett, Michael A. Osborne, Steven Reece and Stephen J. Roberts
Department of Engineering Science
University of Oxford
Oxford, OX1 3PJ
United Kingdom
{rgarnett,mosb,reece,sjrob}@robots.ox.ac.uk

1 Introduction

We introduce a new sequential algorithm for making robust predictions in the presence of changepoints. Unlike many previous approaches [1], which focus on the problem of detecting and locating changepoints, our algorithm focuses on the problem of making predictions even when such changes might be present. We introduce nonstationary covariance functions to be used in Gaussian process prediction that model such changes, then proceed to demonstrate how to effectively manage the hyperparameters associated with those covariance functions. By using Bayesian Monte Carlo, we can integrate out the hyperparameters, allowing us to calculate the marginal predictive distribution. Furthermore, if desired, the posterior distribution over putative changepoint locations can be calculated as a natural byproduct of our prediction algorithm.

2 Gaussian process prediction in the presence of changepoints

Gaussian processes (GPs) offer a powerful method to perform Bayesian inference about functions [2]. A GP is defined as a distribution over the functions $X \to \mathbb{R}$ such that the distribution over the possible function values on any finite subset of X is multivariate Gaussian. For a function $y(x)$, the prior distribution over its values $\boldsymbol{y}$ on a subset $\boldsymbol{x} \subset X$ are completely specified by a mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$, $p(\, \boldsymbol{y} \mid I \,) \triangleq \mathrm{N}(\boldsymbol{y}; \boldsymbol{\mu}(\boldsymbol{x}), \mathbf{K}(\boldsymbol{x}, \boldsymbol{x}))$. Here I, the *context*, includes prior knowledge of both the mean and covariance functions, which generate $\boldsymbol{\mu}$ and $\mathbf{K}$ respectively. We will incorporate knowledge of relevant functional inputs, such as x, into I for notational convenience. The prior mean function is chosen as appropriate for the problem at hand (often a constant), and the covariance function is chosen to reflect any prior knowledge about the structure of the function of interest, for example periodicity.

An example is the squared exponential covariance function, given by $K^{(SE)}(x_1, x_2; \{\lambda, \sigma\}) \triangleq \lambda^2 \exp\left(-\frac{1}{2}(\frac{|x_1 - x_2|}{\sigma})^2\right)$. The parameters λ and σ represent respectively the characteristic *output* and *input scales* of the process. They are examples of the set of hyperparameters, collectively denoted as θ, that are required to specify our covariance and mean functions. Other covariance functions can be constructed for a wide variety of problems [2]. For this reason, GPs are ideally suited for time-series prediction problems with complex behaviour.

In the context of this paper, we will take y to be a potentially dependent dynamic process, such that X contains a time dimension. Note that our approach considers functions of continuous time; we have no need to discretise our observations into time steps. The changepoints we wish to consider will exist only within the term over time. We have developed covariance functions that allow us to model changepoints and faults of many different types [3], some examples of which are displayed in Figure 1. Changepoint covariances are also specified by hyperparameters, such as the location and type of each changepoint.

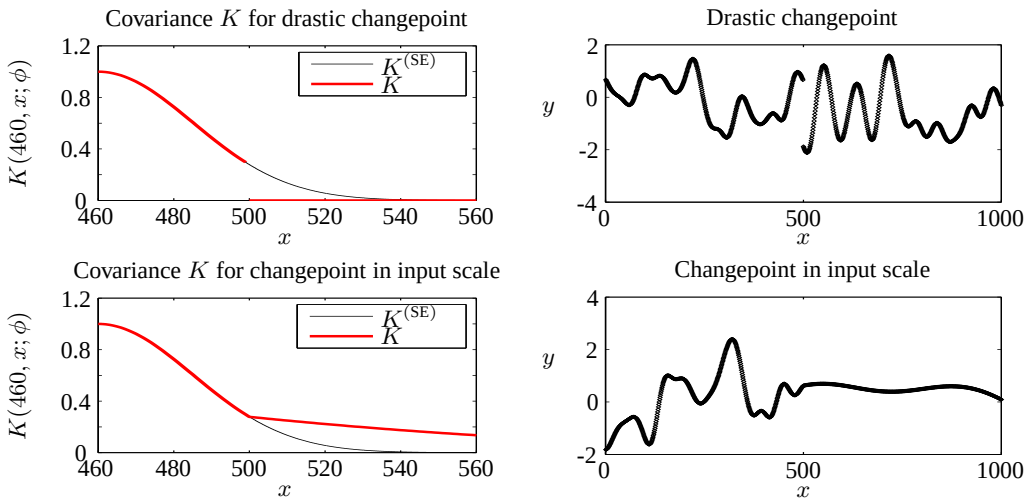

Figure 1: The squared exponential covariance compared with example covariances for the modelling of data with changepoints [3], and associated example data for which they might be appropriate.

Note that we typically do not receive observations of y directly, but rather of potentially corrupted versions z of y. We consider the Gaussian observation likelihood $p(\,\boldsymbol{z} \mid \boldsymbol{y}, \theta, I\,)$. In particular, we often assume independent Gaussian noise contributions of a fixed variance η^2, where η forms part of θ. We define $V(t_1, t_2; \theta) \triangleq K(t_1, t_2; \theta) + \eta^2\,\delta(t_1 - t_2)$. Where we have faults, V must be appropriately modified [3]. A sensor fault essentially implies that the relationship between the underlying, or plant, process y and the observed values z is temporarily altered. As such, $p(\,\boldsymbol{z} \mid \boldsymbol{y}, \theta, I\,)$ will express non-stationary, dependent noise contributions. In order to describe such faults, we include additional hyperparameters into θ specifying their time of occurrence, duration and type. Our principled probabilistic approach will allow us to extract whatever information faulty observations may contain that is pertinent to inference about the plant process.

We define the set of observations available to us as $(\boldsymbol{x}_d, \boldsymbol{z}_d)$. Conditioning on these observations, I, and θ, we are able to analytically derive our predictive equations for the vector of function values $\boldsymbol{y}_\star$ at inputs $\boldsymbol{x}_\star$

$$p(\boldsymbol{y}_\star \mid \boldsymbol{z}_d,\, \theta,\, I) = \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta,\, I)\,, \mathbf{C}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta,\, I)\,\big)\,, \tag{1}$$

where we have

$$\begin{aligned}
\boldsymbol{m}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta,\, I) &\triangleq \boldsymbol{\mu}(\boldsymbol{x}_\star; \theta) + \mathbf{K}(\boldsymbol{x}_\star, \boldsymbol{x}_d; \theta)\mathbf{V}(\boldsymbol{x}_d, \boldsymbol{x}_d; \theta)^{-1}(\boldsymbol{z}_d - \boldsymbol{\mu}(\boldsymbol{x}_d; \theta)) \\
\mathbf{C}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta,\, I) &\triangleq \mathbf{K}(\boldsymbol{x}_\star, \boldsymbol{x}_\star; \theta) - \mathbf{K}(\boldsymbol{x}_\star, \boldsymbol{x}_d; \theta)\mathbf{V}(\boldsymbol{x}_d, \boldsymbol{x}_d; \theta)^{-1}\mathbf{K}(\boldsymbol{x}_d, \boldsymbol{x}_\star; \theta)\,.
\end{aligned}$$

(1) can clearly be used to perform retrospective prediction. Equally, we use the sequential formulation of a GP given by [4] to perform sequential prediction using an adaptive moving window. After each new observation, we use rank-one updates to the covariance matrix to efficiently update our predictions in light of the new information received. We efficiently remove the trailing edge of the window using a similar rank-one "downdate." The computational savings made by these choices mean our algorithm can be feasibly run on-line.

Of course, we can rarely be certain about θ *a priori*. These hyperparameters must hence be assigned an appropriate prior distribution and then marginalized. Although the required integrals are non-analytic, we can efficiently approximate them by use of Bayesian Monte Carlo [5] techniques. This entails evaluating our predictions for a range of hyperparameter samples $\{\theta_i : i \in S\}$, with a different mean $\boldsymbol{m}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)$ and covariance $\mathbf{C}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)$ for each, which are then combined in a weighted mixture

$$\begin{aligned}
p(\,\boldsymbol{y}_\star \mid \boldsymbol{z}_d,\, I\,) &= \frac{\int p(\,\boldsymbol{y}_\star \mid \boldsymbol{z}_d, \theta, I\,)\, p(\,\boldsymbol{z}_d \mid \theta, I\,)\, p(\,\theta \mid I\,)\,\mathrm{d}\theta}{\int p(\,\boldsymbol{z}_d \mid \theta, I\,)\, p(\,\theta \mid I\,)\,\mathrm{d}\theta} \\
&\simeq \sum_{i \in S} \rho_i\, \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)\,, \mathbf{C}(\boldsymbol{y}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)\,\big)\,,
\end{aligned} \tag{2}$$

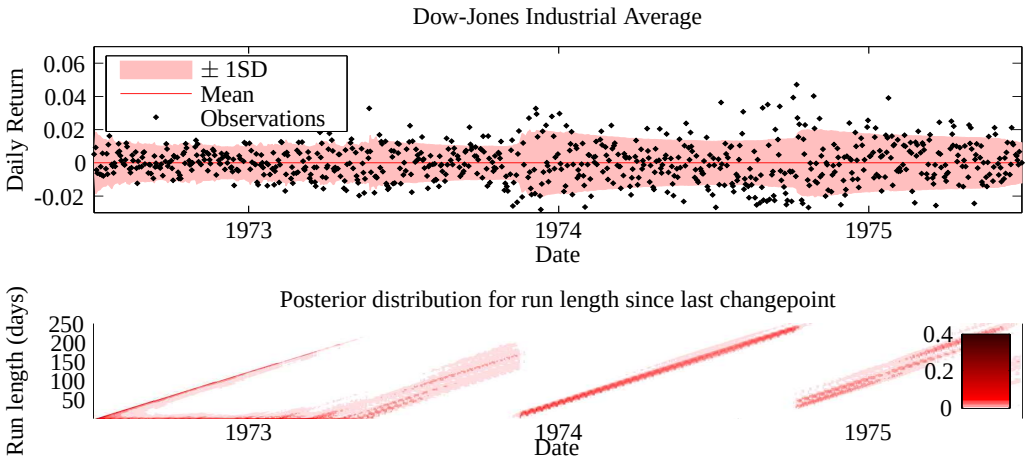

Figure 2: Online predictions and posterior for the location of changepoint for the Dow-Jones data.

with weights $\boldsymbol{\rho}$ as detailed in [4]. To determine the posterior distribution for hyperparameter θ_f by marginalizing over all other hyperparameters θ_{-f}, we must evaluate

$$p(\theta_f \mid \boldsymbol{z}_d,\, I) = \frac{\int p(\boldsymbol{z}_d \mid \theta, I)\, p(\theta \mid I)\, \mathrm{d}\theta_{-f}}{\int p(\boldsymbol{z}_d \mid \theta, I)\, p(\theta \mid I)\, \mathrm{d}\theta} \,. \tag{3}$$

While these integrals are also non-analytic, Bayesian Monte Carlo again gives us a means of approximating them, as elaborated upon in [3].

While the covariance functions from [3] were developed firstly for single changepoints, they are readily extended to handle multiple changepoints. We merely need to introduce further hyperparameters to specify them. In practice, allowing for one or two changepoints within a window is usually sufficient for the purposes of prediction, given that the data prior to a changepoint is typically weakly correlated with data in the current regime of interest. Therefore we can circumvent the computationally onerous task of simultaneously marginalising the hyperparameters associated with the entire data stream. If no changepoint is present in the window, the posterior distribution for its location will typically be concentrated at its trailing edge, where it will have no influence on predictions. Note also that our methods are readily extended to manage changes in mean functions, in addition to changes in covariance functions.

3 Results

We have applied our methods to several datasets. Specifically, we perform prediction for a variable using (2), effectively averaging over models corresponding to a range of changepoints compatible with the data, and also produce a posterior over changepoint/fault locations by estimating (3).

3.1 1972-1975 Dow-Jones industrial average

One canonical changepoint dataset is the series of daily returns of the Dow-Jones industrial average between the 3rd of July, 1972 and the 30th of June, 1975 [6]. We performed sequential prediction on this data using a GP with a diagonal covariance that assumed all measurements were IID. However, our covariance permitted the variance of those observations to undergo changes. Our results are plotted in Figure 2. The *run length* is the number of working days since the last changepoint. Our model clearly identifies important changepoints on the 19th of October, 1973, likely corresponding to the commencement of the OPEC embargo, and the 9th of August, 1974, the date of the resignation of Richard Nixon as President of the USA. A weaker changepoint is identified early in 1973, which [6] speculate is due to the beginning of the Watergate scandal.

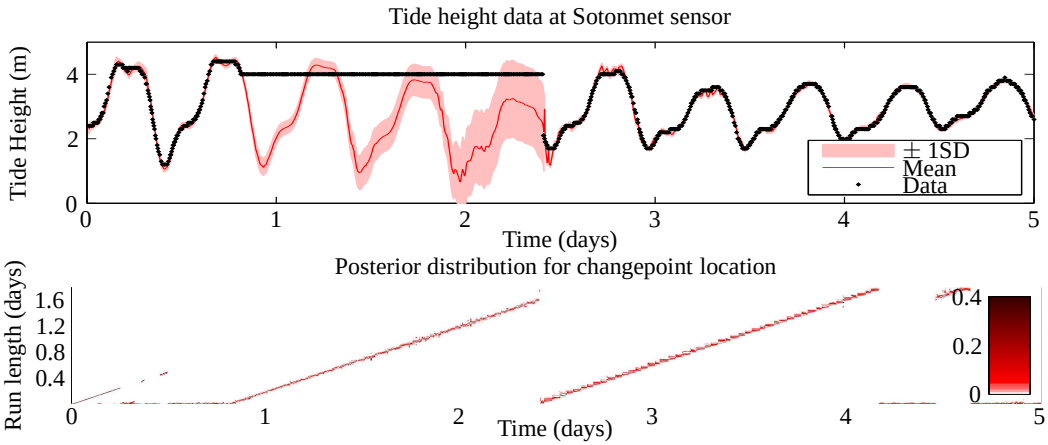

Figure 3: Online predictions and posterior for the location of changepoint for the tide height data.

3.2 Bramblemet weather sensor network

We tested our algorithm on a network of weather sensors located on the south coast of England [4]. In particular, we performed on-line prediction over tide height data in which readings from a sensor became stuck at an incorrect value. As such, we used a model that allowed for such changes in the observation likelihood. Results are plotted in Figure 3. Here, the run length is the number of days since the last changepoint. Our model correctly identified the beginning and end of the fault, and consequently performs effective prediction during its presence.

4 Conclusion

We introduce a new sequential algorithm for performing Bayesian time-series prediction in the presence of changepoints. We use a Gaussian process framework, and develop appropriate covariance functions to model a variety of changepoints. We use Bayesian Monte Carlo numerical integration to estimate the marginal predictive distribution as well as the posterior distribution of associated hyperparameters. By treating the location of a changepoint as a hyperparameter, we may therefore compute the posterior distribution over putative changepoint location as a natural byproduct of our prediction algorithm. Tests on real datasets demonstrate the efficacy of our algorithm.

References

[1] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711, 1995.

[2] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[3] R. Garnett, M. A. Osborne, S. Reece, A. Rogers, and S.J. Roberts. Sequential Bayesian prediction in the presence of changepoints and faults. Technical report, University of Oxford, 2009. available at `http://www.robots.ox.ac.uk/~mosb/PARG0901.pdf`.

[4] M. A. Osborne, A. Rogers, S. Ramchurn, S. J. Roberts, and N. R. Jennings. Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes. In *International Conference on Information Processing in Sensor Networks 2008*, pages 109–120, April 2008.

[5] C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, Cambridge, MA, 2003.

[6] Ryan Prescott Adams and David J.C. MacKay. Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK, 2007. arXiv:0710.3742v1 [stat.ML].