Sequential Bayesian Prediction in the Presence of Changepoints

Roman Garnett, Michael A. Osborne, Stephen J. Roberts
{RGARNETT, MOSB, SJROB}@ROBOTS.OX.AC.UK
Department of Engineering Science, University of Oxford, Oxford, UK OX1 3PJ

Keywords: Gaussian processes, time-series prediction, changepoint detection, Bayesian methods

Abstract

We introduce a new sequential algorithm for making robust predictions in the presence of changepoints. Unlike previous approaches, 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 quadrature, 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.

1. Introduction

We consider the problem of performing time-series prediction in the face of abrupt changes to the properties of the variable of interest. For example, a data stream might undergo a sudden shift in its mean, variance, or characteristic input scale; a periodic signal might have a change in period, amplitude, or phase; or a signal might undergo a change so drastic that its behavior after a particular point in time is completely independent of what happened before. A robust prediction algorithm must be able to make accurate predictions even under such unfavorable conditions.

The problem of detecting and locating abrupt changes in data sequences has been studied under the name *changepoint detection* for decades. A large number of methods have been proposed for this problem; see (Basseville & Nikiforov, 1993; Brodsky & Darkhovsky, 1993; Csorgo & Horvath, 1997; Chen & Gupta, 2000) and the references therein for more information. Relatively few algorithms perform prediction simultaneously with changepoint detection, although sequential Bayesian methods do exist for this problem (Chernoff & Zacks, 1964; Adams & MacKay, 2007). However, these methods, and most methods for changepoint detection in general, make the assumption that the data stream can be segmented into disjoint sequences, such that in each segment the data represent i.i.d. observations from an associated probability distribution. The problem of changepoints in dependent processes has received less attention. Both Bayesian (Carlin et al., 1992; Ray & Tsay, 2002) and non-Bayesian (Muller, 1992; Horváth & Kokoszka, 1997) solutions do exist, although they focus on retrospective changepoint detection alone; their simple dependent models are not employed for the purposes of prediction. Sequential and dependent changepoint detection has been performed (Fearnhead & Liu, 2007) only for a limited set of changepoint models.

We introduce a fully Bayesian framework for performing sequential time-series prediction in the presence of drastic changes in the characteristics of the data. We introduce classes of nonstationary covariance functions to be used in Gaussian process inference for modelling functions with changepoints. In this context, the position of a particular changepoint becomes a hyperparameter of the model. We proceed as usual; for making predictions, the full marginal predictive distribution is estimated. If the locations of changepoints in the data is of interest, we estimate the full posterior distribution of the related hyperparameters given the data. The result is a robust time-series prediction algorithm that makes well-informed predictions even in the presence of sudden changes in the data. If desired, the algo-

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

rithm additionally performs changepoint detection as a natural byproduct of the prediction process.

The remainder of this paper is arranged as follows. In the next section, we briefly introduce Gaussian processes and discuss the marginalization of hyperparameters using Bayesian Monte Carlo numerical integration. A similar technique is presented to produce posterior distributions and their means for any hyperparameters of interest. Next we introduce a class of nonstationary covariance functions to model functions with changepoints. In Section 5 we provide a brief expository example of our algorithm. Finally, we provide results demonstrating the ability of our model to make robust predictions and locate changepoints effectively.

2. Gaussian Process Prediction

Gaussian processes (GPs) offer a powerful method to perform Bayesian inference about functions (Rasmussen & Williams, 2006). 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 set $F \subset X$ is multivariate Gaussian. The prior distribution over the values of a function $y(x)$ are completely specified by a mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$

$$\begin{aligned} p(\,\boldsymbol{y} \mid \boldsymbol{\mu},\, K,\, I\,) &\triangleq \mathrm{N}(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{K}) \\ &\triangleq \frac{1}{\sqrt{\det 2\pi\mathbf{K}}} \exp\left(-\frac{1}{2}\,(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\,\mathbf{K}^{-1}\,(\boldsymbol{y}-\boldsymbol{\mu})\right), \end{aligned}$$

where I, the *context*, includes prior knowledge of both the mean and covariance functions, which generate $\boldsymbol{\mu}$ and $\mathbf{K}$ respectively. 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 or differentiability. A large number of covariance functions exists, and appropriate covariance functions can be constructed for a wide variety of problems (Rasmussen & Williams, 2006). For this reason, GPs are ideally suited for both linear and nonlinear time-series prediction problems with complex behavior. We 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 discretize our observations into time steps.

Our GP distribution is specified by various hyperparameters $\theta_e : e = 1, \ldots, E$, collectively denoted as $\boldsymbol{\theta} \triangleq \{\theta_e : e = 1, \ldots, E\}$. $\boldsymbol{\theta}$ includes the mean function $\boldsymbol{\mu}$, as well as parameters required by the covariance function, input and output scales, amplitudes, periods, etc. as needed.

Define I_d as the conjunction of I and the observations available to us within the window, $(\boldsymbol{x}_d, \boldsymbol{y}_d)$. Taking both I_d and $\boldsymbol{\theta}$ as given, we are able to analytically derive our predictive equations for the vector of function values $\boldsymbol{y}_\star$ at inputs $\boldsymbol{x}_\star$

$$p(\boldsymbol{y}_\star|\boldsymbol{x}_\star,\, \boldsymbol{\theta},\, I_d) = \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y}_\star|I_d), \mathbf{C}_{\boldsymbol{\theta}}(\boldsymbol{y}_\star|I_d)\big)\,, \quad (1)$$

where we have:

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

We use the sequential formulation of a GP given by (Osborne et al., 2008) to perform sequential prediction using a 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.

3. Marginalization

Of course, we can rarely be certain about $\boldsymbol{\theta}$ *a priori*. For each hyperparameter we take an independent Gaussian prior distribution (or if our hyperparameter is restricted to the positive reals, we instead assign a Gaussian distribution to its log) such that

$$p(\,\boldsymbol{\theta} \mid I\,) \triangleq \prod_{e=1}^{E} \mathrm{N}\big(\theta_e; \nu_e, {\lambda_e}^2\big)\,.$$

These hyperparameters must hence be marginalized as

$$\begin{aligned} &p(\boldsymbol{y}_\star|\boldsymbol{x}_\star, I_d) \\ &\quad = \frac{\int p(\boldsymbol{y}_\star|\boldsymbol{x}_\star, \boldsymbol{\theta}, I_d)\, p(\boldsymbol{y}_d|\boldsymbol{x}_d, \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|I)\,\mathrm{d}\boldsymbol{\theta}}{\int p(\boldsymbol{y}_d|\boldsymbol{x}_d, \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|I)\,\mathrm{d}\boldsymbol{\theta}}\,. \end{aligned}$$

Although these required integrals are non-analytic, we can efficiently approximate them by use of Bayesian Monte Carlo (Rasmussen & Ghahramani, 2003) (BMC) techniques. Following (Osborne et al., 2008), we take a grid of hyperparameter samples $\{\boldsymbol{\theta}_s : s = 1, \ldots, S\} \triangleq \boxtimes_{e=1}^{E} \boldsymbol{\Theta}_e$, where $\boldsymbol{\Theta}_e$ is a column vector of samples for the eth hyperparameter and $\boxtimes$ is the Cartesian product. We thus have a different mean $\boldsymbol{m}_s(\boldsymbol{y}_\star|I_d)$, covariance $\mathbf{C}_s(\boldsymbol{y}_\star|I_d)$ and likelihood $l_s \triangleq p(\boldsymbol{y}_d|\boldsymbol{x}_d, \boldsymbol{\theta}_s, I)$ for each. BMC supplies these samples to a GP to perform inference about our integrand

for other values of the hyperparameters. In particular, we assign a *Gaussian* covariance function for this GP

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq \prod_{e=1}^{E} K_e(\theta_e, \theta_e')$$
$$K_e(\theta_e, \theta_e') \triangleq \mathrm{N}(\theta_e; \theta_e', w_e^2)\,.$$

We define

$$\mathfrak{N}_e(\theta_e, \theta_e') \triangleq \mathrm{N}\left(\begin{bmatrix}\theta_e \\ \theta_e'\end{bmatrix}; \begin{bmatrix}\nu_e \\ \nu_e\end{bmatrix}, \begin{bmatrix}\lambda_e^2+w_e^2 & \lambda_e^2 \\ \lambda_e^2 & \lambda_e^2+w_e^2\end{bmatrix}\right)$$
$$\mathfrak{M} \triangleq \bigotimes_{e=1}^{E} \mathbf{K}_e(\boldsymbol{\Theta}_e, \boldsymbol{\Theta}_e)^{-1}\, \mathfrak{N}_e(\boldsymbol{\Theta}_e, \boldsymbol{\Theta}_e)\, \mathbf{K}_e(\boldsymbol{\Theta}_e, \boldsymbol{\Theta}_e)^{-1}$$
$$\boldsymbol{\rho} \triangleq \frac{\mathfrak{M}\, \boldsymbol{l}_S}{\mathbf{1}_{S,1}^{\mathsf{T}}\, \mathfrak{M}\, \boldsymbol{l}_S},$$

where $\mathbf{1}_{S,1}$ is a column vector containing only ones of dimensions equal to $\boldsymbol{l} \triangleq \{l_s : s = 1, \ldots, S\}$, and $\otimes$ is the Kronecker product. Using these, BMC leads us to

$$p(\boldsymbol{y}_\star|\boldsymbol{x}_\star, I_d) \simeq \sum_{s=1}^{S} \rho_s\, \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}_s(\boldsymbol{y}_\star|I_d), \mathbf{C}_s(\boldsymbol{y}_\star|I_d)\big). \tag{2}$$

BMC can also estimate the posterior distribution for hyperparameter θ_f by marginalizing over all other hyperparameters $\boldsymbol{\theta}_{-f}$

$$p(\theta_f|I_d) \quad = \quad \frac{\int p(\boldsymbol{y}_d|\boldsymbol{x}_d, \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta}_{-f}}{\int p(\boldsymbol{y}_d|\boldsymbol{x}_d, \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}|I)\, \mathrm{d}\boldsymbol{\theta}} \quad .$$

With the definitions

$$\mathfrak{K}_{e,f}(\theta_f, \boldsymbol{\Theta}_e) \triangleq \begin{cases} \mathrm{N}(\theta_e; \nu_e, {\lambda_e}^2)\, \mathbf{N}(\boldsymbol{\Theta}_e; \theta_e, w_e^2)^{\mathsf{T}}, & e = f \\ \mathbf{N}(\boldsymbol{\Theta}_e; \nu_e, \lambda_e^2+w_e^2)^{\mathsf{T}}, & e \neq f \end{cases}$$
$$\mathfrak{m}_f^{\mathsf{T}}(\theta_f) \triangleq \bigotimes_{e=1}^{E} \mathfrak{K}_{e,f}(\theta_f, \boldsymbol{\Theta}_e)\, \mathbf{K}_e(\boldsymbol{\Theta}_e, \boldsymbol{\Theta}_e)^{-1},$$
$$\mathfrak{n}^{\mathsf{T}} \triangleq \bigotimes_{e=1}^{E} \mathbf{N}(\boldsymbol{\Theta}_e; \nu_e, \lambda_e^2+w_e^2)^{\mathsf{T}}\, \mathbf{K}_e(\boldsymbol{\Theta}_e, \boldsymbol{\Theta}_e)^{-1}\,,$$

we arrive at

$$p(\theta_f|I_d) \simeq \frac{\mathfrak{m}_f^{\mathsf{T}}(\theta_f)\, \boldsymbol{l}}{\mathfrak{n}^{\mathsf{T}}\, \boldsymbol{l}}\,. \tag{3}$$

Joint posteriors for sets of hyperparameters are also readily obtained in a similar manner. Making the definitions

$$\bar{\mathfrak{K}}_{e,f}(\boldsymbol{\Theta}_e) \triangleq \begin{cases} \frac{\lambda_e^2 \boldsymbol{\Theta}_e^{\mathsf{T}} + w_e^2 \nu_e}{\lambda_e^2+w_e^2} \mathbf{N}(\boldsymbol{\Theta}_e; \nu_e, \lambda_e^2+w_e^2)^{\mathsf{T}}, & e = f \\ \mathbf{N}(\boldsymbol{\Theta}_e; \nu_e, \lambda_e^2+w_e^2)^{\mathsf{T}}, & e \neq f \end{cases}$$
$$\bar{\mathfrak{m}}_f^{\mathsf{T}} \triangleq \bigotimes_{e=1}^{E} \bar{\mathfrak{K}}_{e,f}(\boldsymbol{\Theta}_e)\, \mathbf{K}_e(\boldsymbol{\Theta}_e, \boldsymbol{\Theta}_e)^{-1},$$

the posterior mean is given by

$$\int \theta_f\, p(\theta_f|I_d)\, \mathrm{d}\theta_f \simeq \frac{\bar{\mathfrak{m}}_f^{\mathsf{T}}\, \boldsymbol{l}}{\mathfrak{n}^{\mathsf{T}}\, \boldsymbol{l}}\,. \tag{4}$$

4. Covariance Functions for Prediction in the Presence of Changepoints

We now describe how to construct appropriate covariance functions for functions that experience sudden changes in their characteristics. This section is meant to be expository; the covariance functions we describe are intended as examples rather than an exhaustive list of possibilities. To ease exposition, we assume the input variable of interest x is entirely temporal. If additional features are available, they may be readily incorporated into the derived covariances (Rasmussen & Williams, 2006).

We consider the family of isotropic stationary covariance functions of the form

$$K(x_1, x_2; \{\lambda, \sigma\}) \triangleq \lambda^2 \kappa\left(\tfrac{|x_1-x_2|}{\sigma}\right)\,, \tag{5}$$

where κ is an appropriately chosen function. The parameters λ and σ represent respectively the characteristic *output* and *input scales* of the process. An example isotropic covariance function is the squared exponential covariance, given by

$$K_{SE}(x_1, x_2; \{\lambda, \sigma\}) \triangleq \lambda^2 \exp\left(-\tfrac{1}{2}\left(\tfrac{|x_1-x_2|}{\sigma}\right)^2\right)\,. \tag{6}$$

Many other covariances of the form (5) exist to model functions with a wide range of properties, including the rational quadratic, exponential, and Matérn family of covariance functions. Many choices for κ are also available; for example, to model periodic functions, we can use the covariance

$$K_{PE}(x_1, x_2; \{\lambda, \sigma\}) \triangleq \lambda^2 \exp\left(-\tfrac{1}{2}\sin^2\left(\pi\tfrac{|x_1-x_2|}{\sigma}\right)\right)\,,$$

in which case the output scale λ serves as the amplitude, and the input scale σ serves as the period.

We demonstrate how to construct appropriate covariance functions for three types of changepoints: a sudden change in the input scale, a sudden change in the output scale, and a drastic change rendering values after the changepoint independent of the function values before. The last is the simplest, and we consider it first.

4.1. A drastic change in covariance

Suppose a function of interest is well-behaved except for a drastic change at the point x_c, which separates the function into two regions with associated covariance functions $K_1(\cdot, \cdot; \boldsymbol{\theta}_1)$ before x_c and $K_2(\cdot, \cdot; \boldsymbol{\theta}_2)$ after, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent the values of any hyperparameters associated with K_1 and K_2, respectively. If the change is so drastic that the observations before x_c

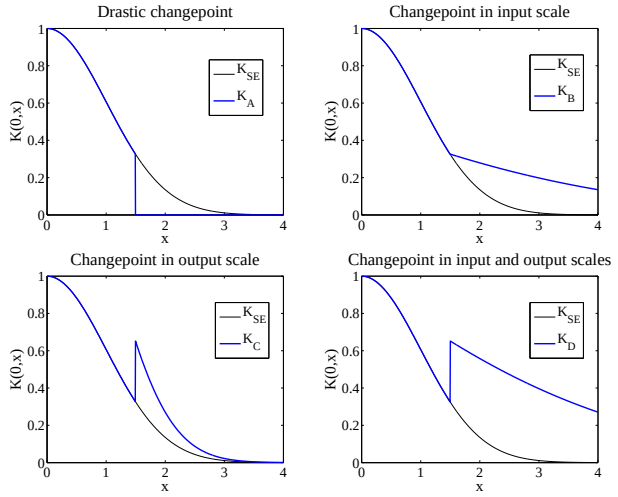

Figure 1. Example covariance functions for the modelling of data with changepoints.

are completely uninformative about the observations after the changepoint; that is, if

$$p\big(\,\boldsymbol{y}_{\geq x_c} \mid I_{<x_c}\,\big) = p\big(\,\boldsymbol{y}_{\geq x_c} \mid I\,\big)\,,$$

where the subscripts indicate ranges of data segmented by x_c, then the appropriate covariance function is trivial. This function can be modelled using the covariance function K_A defined by

$$K_A(x_1, x_2; \boldsymbol{\theta}_A) \triangleq \begin{cases} K_1(x_1, x_2; \boldsymbol{\theta}_1) & (x_1, x_2 < x_c) \\ K_2(x_1, x_2; \boldsymbol{\theta}_2) & (x_1, x_2 \geq x_c) \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The new set of hyperparameters $\boldsymbol{\theta}_A \triangleq \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, x_c\}$ contains knowledge about the original hyperparameters of the covariance functions as well as the location of the changepoint. This covariance function is easily seen to be semi-positive definite and hence admissible.

Theorem 1. *K_A is a valid covariance function.*

Proof. We show that any Gram matrix given by K_A is positive semidefinite. Consider an arbitrary set of input points $\boldsymbol{x}$ in the domain of interest. By appropriately ordering the points in $\boldsymbol{x}$, we may write the Gram matrix $K_A(\boldsymbol{x}, \boldsymbol{x})$ as the block-diagonal matrix

$$\begin{bmatrix} K_1(\boldsymbol{x}_{<x_c}, \boldsymbol{x}_{<x_c}; \boldsymbol{\theta}_1) & \mathbf{0} \\ \mathbf{0} & K_2(\boldsymbol{x}_{\geq x_c}, \boldsymbol{x}_{\geq x_c}; \boldsymbol{\theta}_2) \end{bmatrix};$$

the eigenvalues of $K_A(\boldsymbol{x}, \boldsymbol{x})$ are therefore the eigenvalues of the blocks. Because both K_1 and K_2 are valid covariance functions, their corresponding Gram matrices are positive semidefinite, and therefore eigenvalues of $K_A(\boldsymbol{x}, \boldsymbol{x})$ are nonnegative. □

4.2. A sudden change in input scale

Suppose a function of interest is well-behaved except for a drastic change in the input scale σ at time x_c, which separates the function into two regions with different degrees of long-term dependence.

Let σ_1 and σ_2 represent the input scale of the function before and after the changepoint at x_c, respectively. Suppose we wish to model the function with an isotropic covariance function K of the form (5) that would be appropriate except for the change in input scale. We may model the function using the covariance function K_B defined by

$$K_B(x_1, x_2; \{\lambda^2, \sigma_1, \sigma_2, x_c\}) \triangleq \begin{cases} K(x_1, x_2; \{\lambda, \sigma_1\}) & (x_1, x_2 < x_c) \\ K(x_1, x_2; \{\lambda, \sigma_2\}) & (x_1, x_2 \geq x_c) \\ \lambda^2 \kappa\left(\frac{|x_c - x_1'|}{\sigma_1} + \frac{|x_c - x_2'|}{\sigma_2}\right) & \text{otherwise.} \end{cases} \tag{8}$$

Theorem 2. *K_B is a valid covariance function.*

Proof. Consider the map defined by

$$u(x; x_c) \triangleq \begin{cases} \frac{x}{\sigma_1} & x < x_c \\ \frac{x_c}{\sigma_1} + \frac{x - x_c}{\sigma_2} & x \geq x_c \end{cases}. \tag{9}$$

A simple check shows that $K_B(x_1, x_2; \{\lambda, \sigma_1, \sigma_2, x_c\})$ is equal to $K(u(x_1; x_c), u(x_2; x_c); \{\lambda, 1\})$, the original covariance function with equivalent output scale and unit input scale evaluated on the input points after transformation by u. Because u is injective and K is a valid covariance function, the result follows. □

The function u in the proof above motivates the definition of K_B: by rescaling the input variable appropriately, the change in input scale is removed.

4.3. A sudden change in output scale

Suppose a function of interest is well-behaved except for a drastic change in the output scale λ at time x_c, which separates the function into two regions.

Let $y(x)$ represent the function of interest and let λ_1 and λ_2 represent the output scale of $y(x)$ before and after the changepoint at x_c, respectively. Suppose we wish to model the function with an isotropic covariance function K of the form (5) that would be appropriate except for the change in output scale. To derive the appropriate covariance function, we model $y(x)$ as the product of a function with unit output scale, $g(x)$, and a piecewise-constant scaling function, $a(x)$, defined by

$$a(x; x_c) \triangleq \begin{cases} \lambda_1 & x < x_c \\ \lambda_2 & x \geq x_c \end{cases}. \tag{10}$$

Given the model $y(x) = a(x)g(x)$, the appropriate covariance function for y is immediate. We may use the covariance function K_C defined by

$$K_C(x_1, x_2; \{\lambda_1^2, \lambda_2^2, \sigma, x_c\}) \triangleq a(x_1; x_c)a(x_2; x_c)K(x_1, x_2; \{1, \sigma\}) = \begin{cases} K(x_1, x_2; \{\lambda_1, \sigma\}) & (x_1, x_2 < x_c) \\ K(x_1, x_2; \{\lambda_2, \sigma\}) & (x_1, x_2 \geq x_c) \\ K(x_1, x_2; \{(\lambda_1\lambda_2)^{\frac{1}{2}}, \sigma\}) & \text{otherwise.} \end{cases} \quad (11)$$

The form of K_C follows from the properties of covariance functions, see (Rasmussen & Williams, 2006) for more details.

4.4. Discussion

The key feature of our approach is the treatment of the location and characteristics of changepoints as covariance hyperparameters. As such, for the purposes of prediction, we marginalize them using (2), effectively averaging over models corresponding to a range of changepoints compatible with the data. If desired, the inferred nature of those changepoints can also be directly monitored via (3) and (4).

The covariance functions above can be extended in a number of ways. They can firstly be extended to handle multiple changepoints. Here we need simply to introduce additional hyperparameters for their locations and the values of the appropriate covariance characteristics, such as input scales, within each segment. Note, however, that at any point in time our model only needs to accommodate the volume of data spanned by the window. In practice, allowing for one or two changepoints 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 marginalizing the hyperparameters associated with the entire data stream.

Additionally, if multiple parameters undergo a change at some point in time, an appropriate covariance function can be derived by combining the above results. For example, a function that experiences a change in both input scale and output scale could be readily modeled by

$$K_D(x_1, x_2; \{\lambda_1, \lambda_2, \sigma_1, \sigma_2, x_c\}) \triangleq a(x_1; x_c)a(x_2; x_c)K(u(x_1; x_c), u(x_2; x_c); \{1, 1\}), \quad (12)$$

where u is as defined in (9) and a is as defined in (10).

Notice also that our framework allows for incorporating a possible change in mean, although this does not involve the covariance structure of the model. If the mean function associated with the data is suspected of possible changes, we may treat the mean as a hyperparameter of the model, and place appropriate hyperparameter samples corresponding to, for example, the mean function before and after a putative changepoint. The different possible mean functions will then be properly marginalized for prediction, and the likelihoods associated with the samples can give support for the proposition of a changepoint having occurred at a particular time.

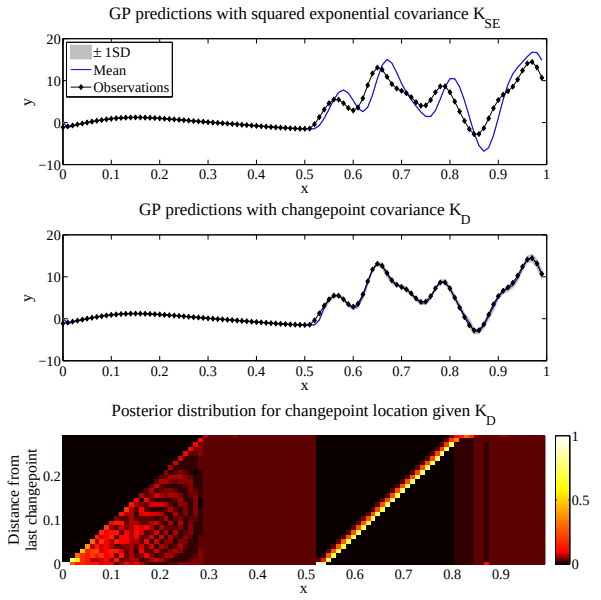

Figure 2. Prediction over a function that undergoes a change in both input scale and output scale.

5. Example

As an expository example, we consider a function that undergoes a sudden change in both input scale and output scale. The function $y(x)$ is displayed in Figure 2; it undergoes a sudden change in input scale (becoming smaller) and output scale (becoming larger) at the point $x = 0.5$. We consider the problem of performing one-step lookahead prediction on $y(x)$ using GP models with a moving window of size 25.

The uppermost plot in Figure 2 shows the performance of a standard GP prediction model with the squared exponential covariance K_{SE} (6), using hyperparameters $\{\lambda, \sigma\}$ selected by maximum likelihood estimation on the data before the changepoint. The standard GP prediction model has clear problems coping with the changepoint; after the changepoint it makes predictions that are very certain (that is, have small predic-

tive variance) that are nonetheless very inaccurate.

The central plot shows the performance of a GP prediction model using the changepoint covariance function K_D (12). The predictions were calculated via BMC hyperparameter marginalization using (2); three samples each were chosen for the hyperparameters $\{\lambda_1, \lambda_2, \sigma_1, \sigma_2\}$, and 25 samples were chosen for the location of the changepoint. Our model easily copes with the changed parameters of the process, continuing to make accurate predictions immediately after the changepoint. Furthermore, by marginalizing the various hyperparameters associated with our model, the uncertainty associated with our predictions is conveyed honestly. The standard deviation becomes roughly an order of magnitude larger after the changepoint due to the similar increase in the output scale.

The lowest plot shows the posterior distribution of the distance to the last changepoint corresponding to the predictions made by the changepoint GP predictor. Each vertical "slice" of the figure at a particular point shows the posterior probability distribution of the distance to the most recent changepoint at that point. The changepoint at $x = 0.5$ is clearly seen in the posterior distribution.

6. Results

6.1. Nile Data

We first consider a canonical changepoint dataset, the minimum water levels of the Nile river during the period AD 622–1284 (Whitcher et al., 2002). Several authors have found evidence supporting a change in input scale for this data around the year AD 722 (Ray & Tsay, 2002). The conjectured reason for this changepoint is the construction in AD 715 of new device (a "nilometer") on the island of Roda, which affected the nature and accuracy of the measurements.

We performed one-step lookahead prediction on this dataset using the input scale changepoint covariance K_B (8), and a moving window of size 100. Seven samples each were used for the hyperparameters σ_1 and σ_2, the input scales before and after a putative changepoint, and fifty samples were used for the location of the changepoint x_c.

The results can be seen in Figure 3. The upper plot shows our predictions for the dataset, including the mean and ± 1 standard deviation error bars. The lower plot shows the posterior distribution of the number of years since the last changepoint. A changepoint around AD 720–722 is clearly visible and agrees with previous results. Several other changepoints are suggested by the posterior distribution; these correspond to locally "rough" patches of the data or very unpredictable points, which suggest a possible constriction in the input scale. Note that the algorithm's confidence in the location of a changepoint does not necessarily correspond with its size; an identified changepoint may represent only a slight shift in input scale.

6.2. EEG Data

We consider EEG data from an epileptic subject (Roberts, 2000). Prediction here is performed with the aim of ultimately building models for EEG activity strong enough to forecast epileptic events. The particular dataset plotted in Figure 4 represents two channels each recorded at 64Hz with 12-bit resolution. It depicts a single epileptic event of the classic "spike and wave" type.

We use a variant of K_A (7) in which the output scales before and after a changepoint may differ. We use also a covariance model (Osborne et al., 2008) that allows us to express the correlation between channels, where the single correlation hyperparameter is also allowed to change at a changepoint. As such, we can model the increased correlation and output scale evident during periods of seizure.

A moving window of size 100 was used to perform the one-step lookahead prediction. Five samples each were used for the hyperparameters λ_1 and λ_2, the output scales before and after a putative changepoint, five samples each were used for the correlation coefficient before and after a putative changepoint, and ten samples were used for the location of the changepoint x_c.

The results can be seen in Figure 4. The upper plot shows our predictions for the dataset, including the mean prediction for each channel and ± 1 standard deviation error bars. The lower plot shows the posterior distribution of the number of seconds since the last changepoint. Several changepoints can be seen in the posterior, including the onset of seizure, as well as changepoints corresponding to each of the individual "spike and wave" events.

Additionally, we show the posterior distributions for the output scale and correlation hyperparameters for the data before and after a putative changepoint at time $t = 5.851$ seconds, as estimated by the model at time $t = 6.125$ seconds. Figure 5 shows the results. The model clearly shows a smaller output scale before the seizure event, and a larger one afterwards. An increase in correlation is also evident, agreeing with expectations about an epileptic event.

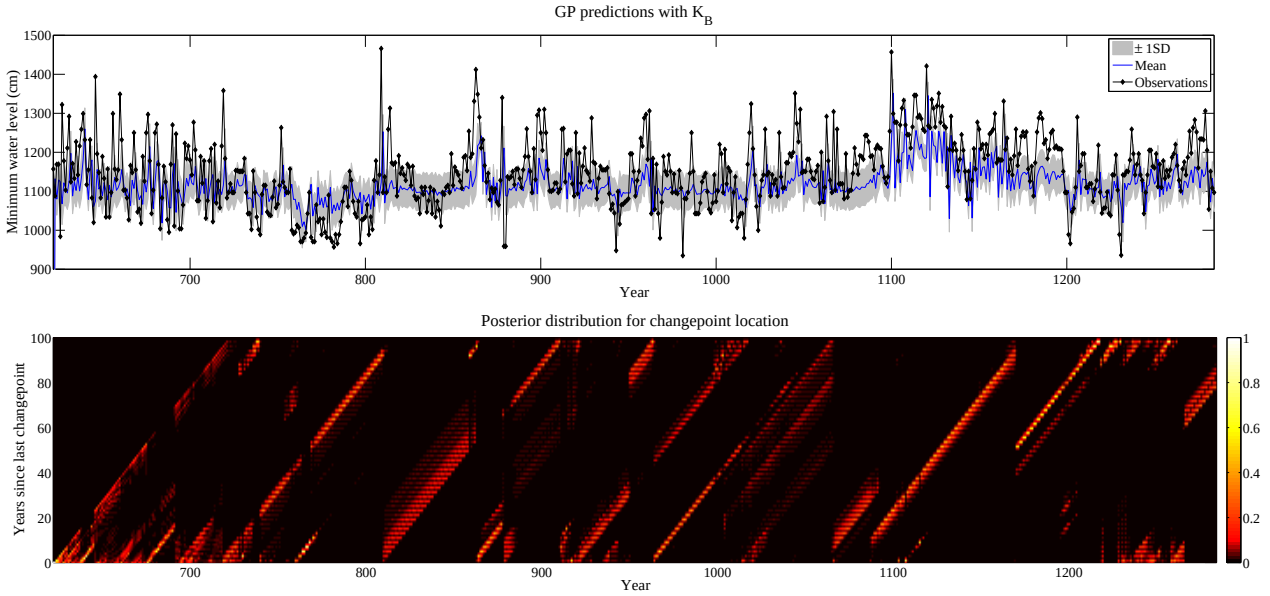

Figure 3. Prediction for the Nile dataset using input scale changepoint covariance K_B, and the corresponding posterior distribution for time since changepoint.

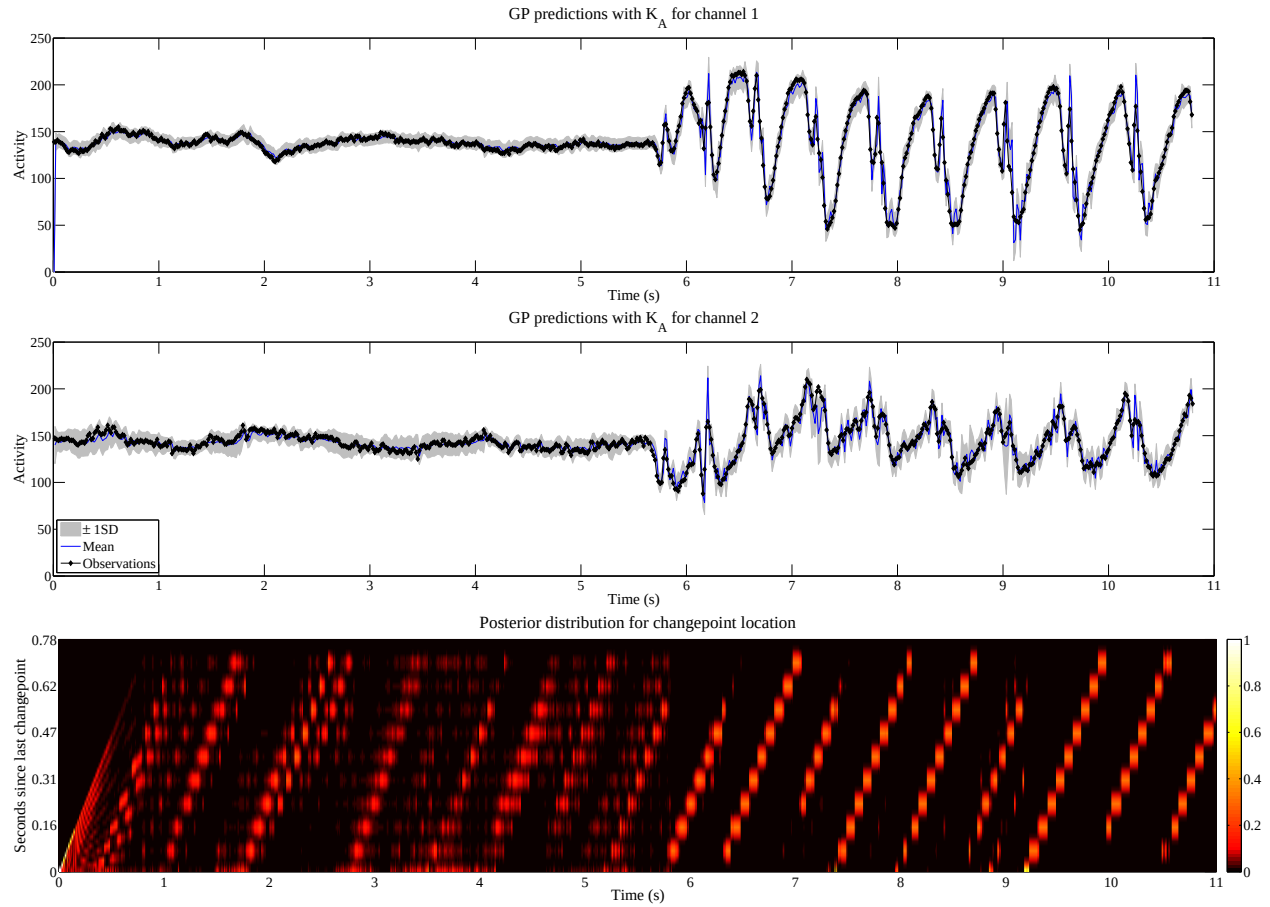

Figure 4. Prediction for the two-channel EEG data using a modified form of K_A, and the corresponding posterior distribution for time since changepoint.

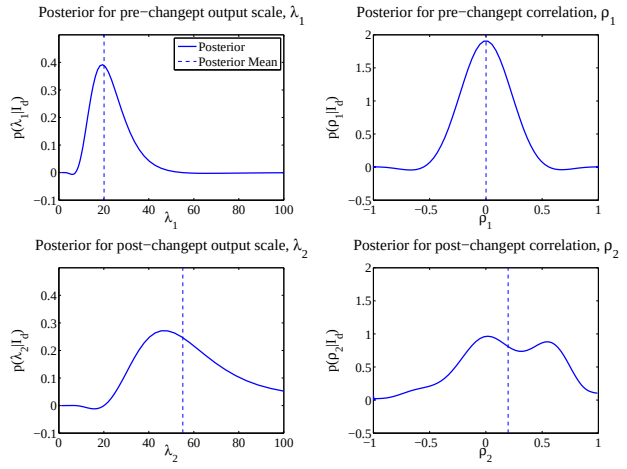

Figure 5. Posteriors and their means for hyperparameters at $t = 6.125$ seconds into the EEG data set (see Figure 4).

7. Conclusion

We introduce a new sequential algorithm for performing Bayesian time-series prediction in the presence of changepoints. After developing appropriate covariance functions to model a variety of changepoints, we incorporate the covariance functions into a Gaussian process framework. 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.

Acknowledgments

This research was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Networks) project and is jointly funded by a BAE Systems and EPSRC strategic partnership (EP/C548051/1).

References

Adams, R. P., & MacKay, D. J. (2007). *Bayesian online changepoint detection* (Technical Report). University of Cambridge, Cambridge, UK. arXiv:0710.3742v1 [stat.ML].

Basseville, M., & Nikiforov, I. (1993). *Detection of abrupt changes: theory and application*. Prentice Hall.

Brodsky, B., & Darkhovsky, B. (1993). *Nonparametric Methods in Change-Point Problems*. Springer.

Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied statistics*, *41*, 389–405.

Chen, J., & Gupta, A. (2000). *Parametric Statistical Change Point Analysis*. Birkháuser Verlag.

Chernoff, H., & Zacks, S. (1964). Estimating the Current Mean of a Normally Distributed Variable Which is Subject to Changes in Time. *Annals of Mathematical Statistics*, *35*, 999–1028.

Csorgo, M., & Horvath, L. (1997). *Limit theorems in change-point analysis*. John Wiley & Sons.

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

Horváth, L., & Kokoszka, P. (1997). The effect of long-range dependence on change-point estimators. *Journal of Statistical Planning and Inference*, *64*, 57–81.

Muller, H. (1992). Change-points in nonparametric regression analysis. *Ann. Statist*, *20*, 737–761.

Osborne, M. A., Rogers, A., Ramchurn, S., Roberts, S. J., & Jennings, N. R. (2008). Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes. *International Conference on Information Processing in Sensor Networks 2008* (pp. 109–120).

Rasmussen, C. E., & Ghahramani, Z. (2003). Bayesian Monte Carlo. In S. Becker and K. Obermayer (Eds.), *Advances in neural information processing systems*, vol. 15. Cambridge, MA: MIT Press.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Ray, B., & Tsay, R. (2002). Bayesian methods for change-point detection in long-range dependent processes. *Journal of Time Series Analysis*, *23*, 687–705.

Roberts, S. J. (2000). Extreme value statistics for novelty detection in biomedical data processing. *Science, Measurement and Technology, IEE Proceedings-* (pp. 363–367).

Whitcher, B., Byers, S., Guttorp, P., & Percival, D. (2002). Testing for homogeneity of variance in time series: Long memory, wavelets and the Nile River. *Water Resources Research*, *38*, 10–1029.