Advance Access publication on February 1, 2010
doi:10.1093/comjnl/bxq003

Sequential Bayesian Prediction in the Presence of Changepoints and Faults

ROMAN GARNETT[1,*] MICHAEL A. OSBORNE[1], STEVEN REECE[1], ALEX ROGERS[2] AND STEPHEN J. ROBERTS[1]

[1]*Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK*
[2]*School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK*
[*]*Corresponding author: rgarnett@robots.ox.ac.uk*

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, and then proceed to demonstrate how to effectively manage the hyperparameters associated with those covariance functions. We further introduce covariance functions to be used in situations where our observation model undergoes changes, as is the case for sensor faults. By using Bayesian quadrature, we can integrate out the hyperparameters, allowing us to calculate the full marginal predictive distribution. Furthermore, if desired, the posterior distribution over putative changepoint locations can be calculated as a natural byproduct of our prediction algorithm.

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

Received 11 August 2009; revised 9 November 2009
Handling editor: Nick Jennings

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 behaviour after a particular point in time is completely independent of what happened before. We also consider cases in which our observations of the variable undergo such changes, even if the variable itself does not; as might occur during a sensor fault. A robust prediction algorithm must be able to make accurate predictions even under such unfavourable 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 [1–4] 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 [5, 6]. 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 [7, 8] and non-Bayesian [9, 10] 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 [11] only for a limited set of changepoint models.

Fault detection, diagnosis and removal is an important application area for sequential time-series prediction in the presence of changepoints. Venkatasubramanian *et al.* [12] classify fault recognition algorithms into three broad categories: quantitative model-based methods, qualitative methods and process history-based methods.

Particularly related to our work are the quantitative methods that employ recursive state estimators. The Kalman filter is commonly used to monitor innovation processes and prediction

error [13, 14]. Banks of Kalman filters have also been applied to fault recognition, where each filter typically corresponds to a specific fault mode [15–17]. Gaussian processes (GPs) are a natural generalization of the Kalman filter and, recently, fault detection has also been studied using GPs [18, 19].

We introduce a fully Bayesian framework for performing sequential time-series prediction in the presence of changepoints. We introduce classes of nonstationary covariance functions to be used in Gaussian process inference for modelling functions with changepoints. We also consider cases in which these changepoints represent a change not in the variable of interest, but instead a change in the function determining our observations of it, as is the case for sensor faults. In such contexts, the position of a particular changepoint becomes a hyperparameter of the model. We proceed as usual when making predictions and evaluate the full marginal predictive distribution. If the locations of changepoints in the data are of interest, we estimate the full posterior distribution of the related hyperparameters conditioned on 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 algorithm additionally performs changepoint and fault 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 GPs and then discuss the marginalization of hyperparameters using Bayesian Monte Carlo numerical integration in Section 3. A similar technique is presented to produce posterior distributions and their means for any hyperparameters of interest. In Section 4, we introduce classes of nonstationary covariance functions to model functions with changepoints or faults. 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. GP PREDICTION

GPs offer a powerful method to perform Bayesian inference about functions [20]. 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. Consider a function $y(x)$. The prior distribution over the values of this function is completely specified by a mean function $\mu(\cdot)$ and a positive-definite covariance function $K(\cdot,\cdot)$. Given these, the distribution of the values of the function at a set of n inputs, $\boldsymbol{x}$, is

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

where I is the *context*, containing all background knowledge pertinent to the problem of inference at hand. We typically incorporate knowledge of relevant functional inputs 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 or a specific amount of differentiability. A large number of covariance functions exist, and appropriate covariance functions can be constructed for a wide variety of problems [20]. For this reason, GPs are ideally suited for both linear and nonlinear 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 discretize our observations into time steps.

Our GP distribution is specified by the values of various hyperparameters collectively denoted θ. These hyperparameters specify the mean function, as well as parameters required by the covariance function: input and output scales, amplitudes, periods etc. as needed.

Note that we typically do not receive observations of y directly, but rather of noise-corrupted versions z of y. We consider only the Gaussian observation likelihood $p(\,\boldsymbol{z} \mid \boldsymbol{y}, \theta, I\,)$. In particular, we typically assume independent Gaussian noise contributions of a fixed variance η^2. This noise variance effectively becomes another hyperparameter of our model and, as such, will be incorporated into θ. To proceed, we define

$$V(x_1, x_2; \theta) \triangleq K(x_1, x_2; \theta) + \eta^2 \delta(x_1 - x_2)\,, \tag{1}$$

where $\delta(\cdot)$ is the Kronecker delta function. Of course, in the noiseless case, $z = y$ and $V(x_1, x_2; \theta) = K(x_1, x_2; \theta)$. We define the set of observations available to us as $(\boldsymbol{x}_d, \boldsymbol{z}_d)$. Taking these observations, I, and θ as given, we are able to analytically derive our predictive equations for the vector of function values $\boldsymbol{y}_\star$ at inputs $\boldsymbol{x}_\star$ as follows:

$$p\big(\boldsymbol{y}_\star \mid \boldsymbol{z}_d, \theta, I\big) = \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}\big(\mathring{\boldsymbol{y}}_\star \mid \boldsymbol{z}_d, \theta, I\big)\,, \mathbf{C}\big(\mathring{\boldsymbol{y}}_\star \mid \boldsymbol{z}_d, \theta, I\big)\big), \tag{2}$$

where we have[1]

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

We also make use of the condensed notation $\boldsymbol{m}_{y|d}(\boldsymbol{x}_\star) \triangleq \boldsymbol{m}\big(\mathring{\boldsymbol{y}}_\star \mid \boldsymbol{y}_d, I\big)$ and $\mathbf{C}_{y|d}(\boldsymbol{x}_\star) \triangleq \mathbf{C}\big(\mathring{\boldsymbol{y}}_\star \mid \boldsymbol{y}_d, I\big)$.

[1] Here the ring accent is used to denote a random variable, e.g. $\mathring{a} = a$ is the proposition that variable $\mathring{a}$ takes the particular value a.

We use the sequential formulation of a GP given by [21] 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 that our algorithm can be feasibly run on-line.

3. MARGINALIZATION

3.1. Posterior predictive distribution

Of course, we can rarely be certain about θ *a priori*, and so we proceed in the Bayesian fashion and marginalize our hyperparameters when necessary.

We assume that our hyperparameter space has finite dimension and write ϕ_e for the value of the eth hyperparameter in θ. We use $\phi_{i,e}$ for the value of the eth hyperparameter in θ_i. For each hyperparameter, we take an independent prior distribution such that

$$p(\theta \mid I) \triangleq \prod_e p(\phi_e \mid I).$$

For any real hyperparameter ϕ_e, we take a Gaussian prior

$$p(\phi_e \mid I) = \mathrm{N}\left(\phi_e; \nu_e, {\lambda_e}^2\right); \tag{3}$$

if our hyperparameter is restricted to the positive reals, we instead assign a Gaussian distribution to its logarithm. For a hyperparameter ϕ_e known only to lie between two bounds l_e and u_e, we take the uniform distribution over that region as follows:

$$p(\phi_e \mid I) = \frac{\sqcap(\phi_e; l_e, u_e)}{\Delta_e}, \tag{4}$$

where $\Delta_e \triangleq u_e - l_e$ and $\sqcap(\theta; l, u)$ is used to denote the rectangular function

$$\sqcap(\phi_e; l_e, u_e) \triangleq \begin{cases} 1, & l_e < \phi_e < u_e \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Occasionally, we may also want to consider a discrete hyperparameter ϕ_e. In this case, we take the uniform prior

$$P(\phi_e \mid I) = \frac{1}{\Delta_e}, \tag{6}$$

where Δ_e is here defined as the number of discrete values the hyperparameter can take.

Our hyperparameters must then be marginalized as

$$p(\boldsymbol{y}_\star \mid z_d, I) = \frac{\int p(\boldsymbol{y}_\star \mid z_d, \theta, I)\, p(z_d \mid \theta, I)\, p(\theta \mid I)\, \mathrm{d}\theta}{\int p(z_d \mid \theta, I)\, p(\theta \mid I)\, \mathrm{d}\theta}. \tag{7}$$

Although these required integrals are non-analytic, we can efficiently approximate them by the use of Bayesian Quadrature (BQ) [22] techniques. As with any method of quadrature, we require a set of samples of our integrand. Following [21], we take a grid of hyperparameter samples $\boldsymbol{\theta}_s \triangleq \times_e \boldsymbol{\phi}_{u,e}$, where $\boldsymbol{\phi}_{u,e}$ is a column vector of unique samples for the eth hyperparameter and $\times$ is the Cartesian product. We thus have a different mean, covariance and likelihood for each sample. Of course, this sampling is necessarily sparse in hyperparameter space. For θ far from our samples, $\boldsymbol{\theta}_s$, we are uncertain about the values of the two terms in our integrand: the predictions

$$\mathring{q}(\theta) \triangleq p(\boldsymbol{y}_\star \mid z_d, \theta, I),$$

and the likelihoods

$$\mathring{r}(\theta) \triangleq p(z_d \mid \theta, I).$$

It is important to note that the function q evaluated at a point θ returns a *function* (a predictive distribution for $\boldsymbol{y}_\star$), whereas the function r evaluated at a point θ returns a *scalar* (a marginal likelihood).

To estimate (6), BQ begins by assigning GP priors to both q and r. Given our (noiseless) observations of these functions, $\boldsymbol{q}_s \triangleq q(\boldsymbol{\theta}_s)$ and $\boldsymbol{r}_s \triangleq r(\boldsymbol{\theta}_s)$, the GPs allow us to perform inference about the function values at any other point. Because integration is a projection, and variables over which we have a multivariate Gaussian distribution are joint Gaussian with any affine transformation of those variables, our GP priors then allow us to use our samples of the integrand to perform an inference about the integrals. We define our unknown variables

$$\mathring{\varrho} \triangleq p(\boldsymbol{y}_\star \mid z_d, I) = \frac{\int q(\theta) r(\theta) p(\theta \mid I)\, \mathrm{d}\theta}{\int r(\theta) p(\theta \mid I)\, \mathrm{d}\theta}.$$

and

$$m(\mathring{\varrho} \mid \boldsymbol{q}_s, r, I) \triangleq \frac{\int m_{q|s}(\theta) r(\theta) p(\theta \mid I)\, \mathrm{d}\theta}{\int r(\theta) p(\theta \mid I)\, \mathrm{d}\theta},$$

in order to proceed as follows:

$$\begin{aligned} p(\boldsymbol{y}_\star \mid \boldsymbol{q}_s, \boldsymbol{r}_s, z_d, I) &= \iiint p(\boldsymbol{y}_\star \mid q, r, z_d, I)\, p(\varrho \mid q, r, I) \\ &\quad \times p(q \mid \boldsymbol{q}_s, I)\, p(r \mid \boldsymbol{r}_s, I)\, \mathrm{d}\varrho \mathrm{d}q \mathrm{d}r \\ &= \iiint \varrho \delta(\varrho - \mathring{\varrho})\, \mathrm{N}(q; m_{q|s}, C_{q|s}) \\ &\quad \times \mathrm{N}(r; m_{r|s}, C_{r|s})\, \mathrm{d}\varrho \mathrm{d}q \mathrm{d}r \\ &= \int m(\mathring{\varrho} \mid \boldsymbol{q}_s, r, I)\, \mathrm{N}(r; m_{r|s}, C_{r|s})\, \mathrm{d}r. \end{aligned}$$

Here our integration again becomes non-analytic. As a consequence, we take a maximum *a posteriori* (MAP) approximation for r, which approximates $\mathrm{N}(r; m_{r|s}, C_{r|s})$ as

$\delta(r - m_{r|s})$. This gives us

$$p(\boldsymbol{y}_\star \mid \boldsymbol{q}_s, \boldsymbol{r}_s, \boldsymbol{z}_d, I) \overset{\propto}{\sim} \int m_{q|s}(\theta) m_{r|s}(\theta) p(\theta \mid I)\, \mathrm{d}\theta.$$

We now take the independent product Gaussian covariance function for our GPs over both q and r as follows:

$$\begin{aligned} K(\theta_i, \theta_j) &\triangleq \prod_e K_e(\phi_{i,e}, \phi_{j,e}) \\ K_e(\phi_{i,e}, \phi_{j,e}) &\triangleq \mathrm{N}(\phi_{i,e}; \phi_{j,e}, w_e^2), \end{aligned} \tag{8}$$

and so, defining

$$\begin{aligned} \mathfrak{N}_e(\phi_{i,e}, \phi_{j,e}) \triangleq \int & K_e(\phi_{i,e}, \phi_{*,e}) p(\phi_{*,e} \mid I) \\ & \times K_e(\phi_{*,e}, \phi_{j,e}) \mathrm{d}\phi_{*,e}, \end{aligned}$$

we have

$$\mathfrak{N}_e(\phi_{i,e}, \phi_{j,e}) = \mathrm{N}\left(\begin{bmatrix} \phi_{i,e} \\ \phi_{j,e} \end{bmatrix}; \begin{bmatrix} \nu_e \\ \nu_e \end{bmatrix}, \begin{bmatrix} \lambda_e^2 + w_e^2 & \lambda_e^2 \\ \lambda_e^2 & \lambda_e^2 + w_e^2 \end{bmatrix} \right),$$

if $p(\phi_e \mid I)$ is the Gaussian (3), and

$$\begin{aligned} \mathfrak{N}_e(\phi_{i,e}, \phi_{j,e}) = \mathrm{N}(\phi_{i,e}; \phi_{j,e}, 2\, w_e^2) & \\ \times \left(\Phi\left(u_e; \frac{1}{2}(\phi_{i,e} + \phi_{j,e}), \frac{1}{2} w_e^2 \right) \right. & \\ \left. - \Phi\left(l_e; \frac{1}{2}(\phi_{i,e} + \phi_{j,e}), \frac{1}{2} w_e^2 \right) \right), & \end{aligned}$$

if $p(\phi_e \mid I)$ is the uniform (4). We use Φ to represent the usual Gaussian cumulative distribution function. Finally, we have

$$\mathfrak{N}_e(\phi_{i,e}, \phi_{j,e}) = \sum_{d=1}^{\Delta_e} \frac{1}{\Delta_e} K_e(\phi_{i,e}, \phi_{d,e}) K_e(\phi_{d,e}, \phi_{j,e}),$$

if $p(\phi_e \mid I)$ is the discrete uniform (6). We now make the further definitions

$$\mathfrak{M} \triangleq \bigotimes_e \mathbf{K}_e(\boldsymbol{\phi}_{u,e}, \boldsymbol{\phi}_{u,e})^{-1} \mathfrak{N}_e(\boldsymbol{\phi}_{u,e}, \boldsymbol{\phi}_{u,e}) \mathbf{K}_e(\boldsymbol{\phi}_{u,e}, \boldsymbol{\phi}_{u,e})^{-1}$$

$$\boldsymbol{\gamma} \triangleq \frac{\mathfrak{M} \boldsymbol{r}_s}{\mathbf{1}_s^\mathsf{T} \mathfrak{M} \boldsymbol{r}_s}, \tag{9}$$

where $\mathbf{1}_s$ is a column vector containing only ones of dimensions equal to $\boldsymbol{r}_s$, and $\otimes$ is the Kronecker product. Using these, BQ leads us to

$$\begin{aligned} & p(\boldsymbol{y}_\star \mid \boldsymbol{q}_s, \boldsymbol{r}_s, \boldsymbol{z}_d, I) \\ & \quad \simeq \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{q}_s \\ & \quad = \sum_i \gamma_i \mathrm{N}(\boldsymbol{y}_\star; \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \mid \boldsymbol{z}_d, \theta_i, I), \mathbf{C}(\mathring{\boldsymbol{y}}_\star \mid \boldsymbol{z}_d, \theta_i, I)). \end{aligned} \tag{10}$$

That is, our final posterior is a weighted mixture of the Gaussian predictions produced by each hyperparameter sample. This is the reason for the form of (8)—we know that $p(\boldsymbol{y}_\star \mid \boldsymbol{z}_d, I)$ must integrate to one, and therefore $\sum_i \gamma_i = 1$.

3.2. Hyperparameter posterior distribution

We can also use BQ to estimate the posterior distribution for hyperparameter ϕ_f (which could, in general, also represent a set of hyperparameters) by marginalizing over all other hyperparameters ϕ_{-f}

$$p(\phi_f \mid \boldsymbol{z}_d, I) = \frac{\int p(\boldsymbol{z}_d \mid \theta, I)\, p(\theta \mid I)\, \mathrm{d}\phi_{-f}}{\int p(\boldsymbol{z}_d \mid \theta, I)\, p(\theta \mid I)\, \mathrm{d}\theta}.$$

Here we can again take a GP for r and use it to perform an inference about $\mathring{\rho} \triangleq p(\phi_f \mid \boldsymbol{z}_d, I)$. We define

$$m(\mathring{\rho} \mid r, I) = \frac{\int r(\theta) p(\theta \mid I)\, \mathrm{d}\phi_{-f}}{\int r(\theta) p(\theta \mid I)\, \mathrm{d}\theta},$$

and can then write

$$\begin{aligned} & p(\phi_f \mid \boldsymbol{r}_s, \boldsymbol{z}_d, I) \\ & \quad = \int p(\phi_f \mid r, \boldsymbol{z}_d, I)\, p(\rho \mid r, I)\, p(r \mid \boldsymbol{r}_s, I)\, \mathrm{d}\rho \mathrm{d}r \\ & \quad = \int \rho \delta(\rho - m(\mathring{\rho} \mid r, I)) \mathrm{N}(r; m_{r|s}, C_{r|s})\, \mathrm{d}\rho \mathrm{d}r. \end{aligned}$$

As before, we take a MAP approximation for r to give us

$$p(\phi_f \mid \boldsymbol{r}_s, \boldsymbol{z}_d, I) \overset{\propto}{\sim} \int m_{r|s}(\theta)\, p(\theta \mid I)\, \mathrm{d}\phi_{-f}.$$

We again take the covariance defined by (7), and define

$$\mathfrak{K}_{e,f}(\phi_e, \phi_{i,e}) \triangleq \begin{cases} K_e(\phi_e, \phi_{i,e}) p(\phi_e \mid I), & e \in f \\ \int K_e(\phi_e, \phi_{i,e}) p(\phi_e \mid I)\, \mathrm{d}\phi_e, & e \notin f \end{cases},$$

which leads to

$$\mathfrak{K}_{e,f}(\phi_e, \phi_{i,e}) = \begin{cases} \mathrm{N}(\phi_e; \phi_{i,e}, w_e^2)\, \mathrm{N}(\phi_e; \nu_e, {\lambda_e}^2), & e \in f \\ \mathrm{N}(\phi_{i,e}; \nu_e, \lambda_e^2 + w_e^2), & e \notin f \end{cases},$$

if $p(\phi_e \mid I)$ is the Gaussian (3);

$$\begin{aligned} & \mathfrak{K}_{e,f}(\phi_e, \phi_{i,e}) \\ & \quad = \begin{cases} \mathrm{N}(\phi_e; \phi_{i,e}, w_e^2) \dfrac{\sqcap(\phi_e; l_e, u_e)}{\Delta_e}, & e \in f \\ \dfrac{1}{\Delta_e} \left(\Phi\left(u_e; \phi_{i,e}, w_e^2\right) - \Phi\left(l_e; \phi_{i,e}, w_e^2\right) \right), & e \notin f \end{cases}, \end{aligned}$$

if $p(\phi_e \mid I)$ is the uniform (4); and

$$\mathfrak{K}_{e,f}(\phi_e, \phi_{i,e}) = \begin{cases} \dfrac{1}{\Delta_e} K_e(\phi_e, \phi_{i,e}), & e \in f \\ \sum_{d=1}^{\Delta_e} \dfrac{1}{\Delta_e} K_e(\phi_{d,e}, \phi_{i,e}), & e \notin f \end{cases},$$

if $p(\phi_e \mid I)$ is the discrete uniform (5). We now define

$$\mathfrak{m}_f^\mathsf{T}(\phi_f) \triangleq \bigotimes_e \mathfrak{K}_{e,f}(\phi_f, \boldsymbol{\phi}_{u,e})^\mathsf{T} \mathbf{K}_e(\boldsymbol{\phi}_{u,e}, \boldsymbol{\phi}_{u,e})^{-1},$$

and arrive at

$$p\left(\phi_f \mid \boldsymbol{r}_s, \boldsymbol{z}_d, I\right) \simeq \frac{\mathfrak{m}_f^{\mathsf{T}}(\phi_f)\boldsymbol{r}_s}{\mathfrak{m}_\emptyset^{\mathsf{T}}(\emptyset)\boldsymbol{r}_s}, \tag{11}$$

where $\emptyset$ is the empty set; for $\mathfrak{m}_\emptyset^{\mathsf{T}}(\emptyset)$ we use the definitions above with $f = \emptyset$. This factor will ensure the correct normalization of our posterior.

3.3. Hyperparameter posterior mean

For a more precise idea about our hyperparameters, we can use BQ one final time to estimate the posterior mean for a hyperparameter ϕ_f

$$m\left(\mathring{\phi}_f \mid \boldsymbol{z}_d, I\right) = \frac{\int \phi_f \; p(\boldsymbol{z}_d \mid \theta, I)\; p(\theta \mid I)\,\mathrm{d}\theta}{\int p(\boldsymbol{z}_d \mid \theta, I)\; p(\theta \mid I)\,\mathrm{d}\theta}.$$

Essentially, we take exactly the same approach as in Section 3.2. Making the definition

$$\bar{\mathfrak{K}}_{e,f}(\phi_{i,e}) \triangleq \begin{cases} \int \phi_e \, K_e(\phi_e, \phi_{i,e}) p(\phi_e \mid I)\,\mathrm{d}\phi_e, & e \in f \\ \int K_e(\phi_e, \phi_{i,e}) p(\phi_e \mid I)\,\mathrm{d}\phi_e, & e \notin f \end{cases},$$

we arrive at

$$\bar{\mathfrak{K}}_{e,f}(\phi_{i,e}) = \begin{cases} \mathrm{N}\left(\phi_{i,e}; \nu_e, \lambda_e^2 + w_e^2\right) \dfrac{\lambda_e^2 \phi_{i,e} + w_e^2 \nu_e}{\lambda_e^2 + w_e^2}, & e \in f \\ \mathrm{N}\left(\phi_{i,e}; \nu_e, \lambda_e^2 + w_e^2\right), & e \notin f \end{cases},$$

if $p(\phi_e \mid I)$ is the Gaussian (3);

$$\bar{\mathfrak{K}}_{e,f}(\phi_{i,e}) = \begin{cases} \dfrac{\phi_{i,e}}{\Delta_e}\left(\Phi\left(u_e; \phi_{i,e}, w_e^2\right) - \Phi\left(l_e; \phi_{i,e}, w_e^2\right)\right) \\ \quad - \dfrac{w_e^2}{\Delta_e}\left(\mathrm{N}\left(u_e; \phi_{i,e}, w_e^2\right) - \mathrm{N}\left(l_e; \phi_{i,e}, w_e^2\right)\right), & e \in f \\ \dfrac{1}{\Delta_e}\left(\Phi\left(u_e; \phi_{i,e}, w_e^2\right) - \Phi\left(l_e; \phi_{i,e}, w_e^2\right)\right), & e \notin f \end{cases},$$

if $p(\phi_e \mid I)$ is the uniform (4); and

$$\bar{\mathfrak{K}}_{e,f}(\phi_{i,e}) = \begin{cases} \sum_{d=1}^{\Delta_e} \dfrac{\phi_{d,e}}{\Delta_e} K_e(\phi_{d,e}, \phi_{i,e}), & e \in f \\ \sum_{d=1}^{\Delta_e} \dfrac{1}{\Delta_e} K_e(\phi_{d,e}, \phi_{i,e}), & e \notin f \end{cases},$$

if $p(\phi_e \mid I)$ is the discrete uniform (6). We now make the corresponding definition

$$\bar{\mathfrak{m}}_f^{\mathsf{T}} \triangleq \bigotimes_e \bar{\mathfrak{K}}_{e,f}\left(\boldsymbol{\phi}_{u,e}\right)^{\mathsf{T}} \mathbf{K}_e\left(\boldsymbol{\phi}_{u,e}, \boldsymbol{\phi}_{u,e}\right)^{-1},$$

giving the posterior mean as

$$m\left(\mathring{\phi}_f \mid \boldsymbol{z}_d, I\right) \simeq \frac{\bar{\mathfrak{m}}_f^{\mathsf{T}} \boldsymbol{r}_s}{\bar{\mathfrak{m}}_\emptyset^{\mathsf{T}} \boldsymbol{r}_s}. \tag{12}$$

Note that $\bar{\mathfrak{m}}_\emptyset^{\mathsf{T}} = \mathfrak{m}_\emptyset^{\mathsf{T}}(\emptyset)$.

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 that the input variable of interest x is entirely temporal. If additional features are available, they may be readily incorporated into the derived covariances [20].

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

$$K(x_1, x_2; \{\lambda, \sigma\}) \triangleq \lambda^2 \kappa\left(\frac{|x_1 - x_2|}{\sigma}\right), \tag{13}$$

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

$$K_{SE}(x_1, x_2; \{\lambda, \sigma\}) \triangleq \lambda^2 \exp\left(-\frac{1}{2}\left(\frac{|x_1 - x_2|}{\sigma}\right)^2\right). \tag{14}$$

Many other covariances of the form (13) 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

$$K_{PE}(x_1, x_2; \{\lambda, \sigma\}) \triangleq \lambda^2 \exp\left(-\frac{1}{2\omega}\sin^2\left(\pi\frac{|x_1 - x_2|}{\sigma}\right)\right), \tag{15}$$

in which case the output scale λ serves as the amplitude, and the input scale σ serves as the period. We have ω as a roughness parameter that serves a role similar to the input scale σ in (13).

We now demonstrate how to construct appropriate covariance functions for a number of types of changepoint. Some examples of these are illustrated in Fig. 1.

4.1. A drastic change in covariance

Suppose that 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 $K_1(\cdot, \cdot; \theta_1)$ before x_c and $K_2(\cdot, \cdot; \theta_2)$ after, where θ_1 and θ_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 are completely uninformative about the observations after the changepoint; that is, if

$$p\left(\boldsymbol{y}_{\geq x_c} \mid \boldsymbol{z}, I\right) = p\left(\boldsymbol{y}_{\geq x_c} \mid \boldsymbol{z}_{\geq x_c}, I\right),$$

where the subscripts indicate ranges of data segmented by x_c (e.g. $\boldsymbol{z}_{\geq x_c}$ is the subset of $\boldsymbol{z}$ containing only observations after the

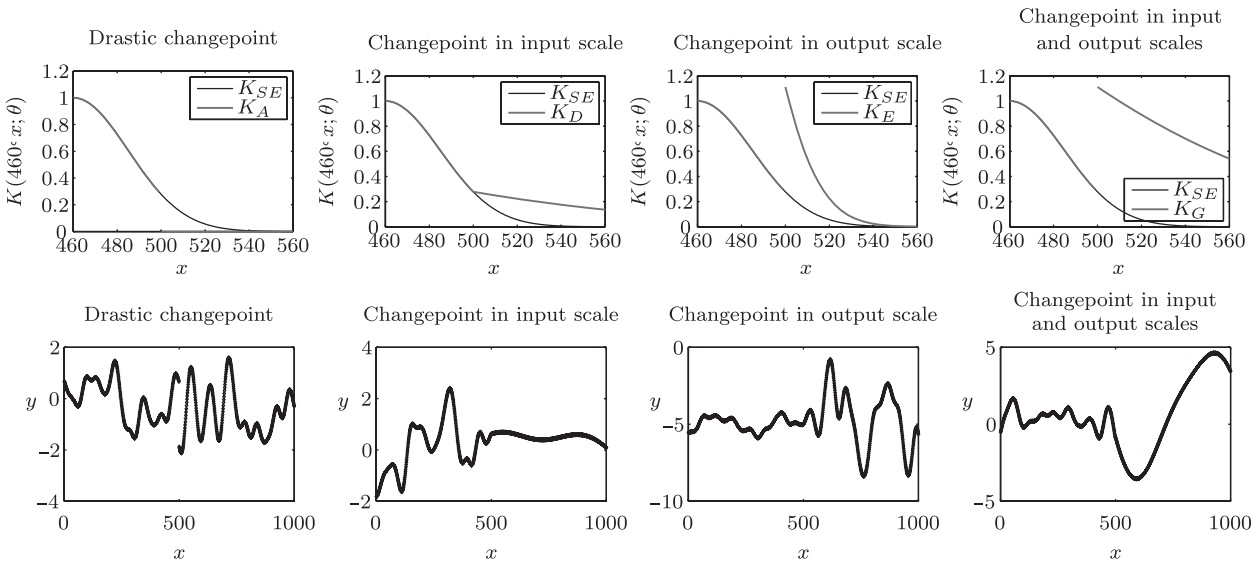

FIGURE 1. Example covariance functions for the modelling of data with changepoints, and associated example data for which they might be appropriate.

changepoint), then the appropriate covariance function is trivial. This function can be modelled using the covariance function K_A defined by

$$K_A(x_1, x_2; \theta_A) \triangleq \begin{cases} K_1(x_1, x_2; \theta_1), & x_1, x_2 < x_c \\ K_2(x_1, x_2; \theta_2), & x_1, x_2 \geq x_c \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

The new set of hyperparameters $\theta_A \triangleq \{\theta_1, \theta_2, x_c\}$ 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 4.1. *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

$$\begin{bmatrix} K_1(\boldsymbol{x}_{<x_c}, \boldsymbol{x}_{<x_c}; \theta_1) & \mathbf{0} \\ \mathbf{0} & K_2(\boldsymbol{x}_{\geq x_c}, \boldsymbol{x}_{\geq x_c}; \theta_2) \end{bmatrix};$$

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 non-negative. □

4.2. A smooth drastic change in covariance

Suppose that a *continuous function* of interest is best modelled by different covariance functions, before and after a changepoint x_c. The function values after the changepoint are conditionally independent of the function values before, given the value at the changepoint itself. The Bayesian network for this probabilistic structure is depicted in Fig. 2. This represents an extension to the drastic covariance described above; our two regions can be drastically different, but we can still enforce smoothness across the boundary between them.

The changepoint separates the function into two regions with associated covariance functions $K_1(\cdot, \cdot; \theta_1)$ before x_c and $K_2(\cdot, \cdot; \theta_2)$ after, where θ_1 and θ_2 represent the values of any hyperparameters associated with K_1 and K_2, respectively. We introduce a further hyperparameter, k_c, which represents the covariance function value at the changepoint. We may model the function using the covariance function K_B defined by

$$K_B(x_1, x_2; \theta_1, \theta_2) \triangleq \begin{cases} K_1(x_1, x_2; \theta_1) + G_1(k_c - K_1(x_c, x_c; \theta_1))G_1^{\mathsf{T}}, \\ \quad x_1, x_2 < x_c \\ K_2(x_1, x_2; \theta_2) + G_2(k_c - K_2(x_c, x_c; \theta_2))G_2^{\mathsf{T}}, \\ \quad x_1, x_2 > x_c \\ G_1 k_c G_2^T, \quad \text{otherwise} \end{cases} \quad (17)$$

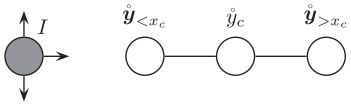

FIGURE 2. Bayesian Network for the smooth drastic change model. I is the context, correlated with all other nodes.

where

$$G_1 = \frac{K_1(x_1, x_c; \theta_1)}{K_1(x_c, x_c; \theta_1)} \quad \text{and} \quad G_2 = \frac{K_2(x_2, x_c; \theta_2)}{K_2(x_c, x_c; \theta_2)}.$$

We call this covariance function the *continuous conditionally independent* covariance function. This covariance function can be extended to multiple changepoints, boundaries in multi-dimensional spaces, and also to cases where function derivatives are continuous at the changepoint. For proofs and details of this covariance function the reader is invited to see [19].

As a slight extension of this covariance, consider a function that undergoes a temporary excursion from an otherwise constant value of zero. This excursion is known to be smooth, that is, it both begins and ends at zero. We define the beginning of the excursion as x_{c_1} and its end as x_{c_2}. Essentially, we have changepoints as considered by (17) at both x_{c_1} and x_{c_2}. We can hence write the covariance function appropriate for this function as

$$\begin{aligned} &K_C(x_1, x_2; \{\theta, x_{c_1}, x_{c_2}\}) \\ &\quad \triangleq K(x_1, x_2; \theta) - K\left(x_1, \begin{bmatrix} x_{c_1} \\ x_{c_2} \end{bmatrix}; \theta\right) K\left(\begin{bmatrix} x_{c_1} \\ x_{c_2} \end{bmatrix}, \begin{bmatrix} x_{c_1} \\ x_{c_2} \end{bmatrix}; \theta\right)^{-1} \\ &\qquad \times K\left(\begin{bmatrix} x_{c_1} \\ x_{c_2} \end{bmatrix}, x_2; \theta\right), \end{aligned} \tag{18}$$

for $x_{c_1} < x_1 < x_{c_2}$ and $x_{c_1} < x_2 < x_{c_2}$, and $K_C(x_1, x_2; \{\theta, x_{c_1}, x_{c_2}\}) = 0$ otherwise. Here (unsubscripted) K is a covariance function that describes the dynamics of the excursion itself.

4.3. A sudden change in input scale

Suppose that 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 that we wish to model the function with an isotropic covariance function K of the form (12) that would be appropriate except for the change in input scale. We may model the function using the covariance function K_D defined by

$$\begin{aligned} &K_D(x_1, x_2; \{\lambda^2, \sigma_1, \sigma_2, x_c\}) \\ &\quad \triangleq \begin{cases} K(x_1, x_2; \{\lambda, \sigma_1\}), & x_1, x_2 < x_c \\ K(x_1, x_2; \{\lambda, \sigma_2\}), & x_1, x_2 \geq x_c \\ \lambda^2 \kappa\left(\dfrac{|x_c - x_1'|}{\sigma_1} + \dfrac{|x_c - x_2'|}{\sigma_2}\right), & \text{otherwise} \end{cases}. \end{aligned} \tag{19}$$

THEOREM 4.2. *We have that K_D is a valid covariance function.*

Proof. Consider the map defined by

$$u(x; x_c) \triangleq \begin{cases} \dfrac{x}{\sigma_1}, & x < x_c \\ \dfrac{x_c}{\sigma_1} + \dfrac{x - x_c}{\sigma_2}, & x \geq x_c \end{cases}. \tag{20}$$

A simple check shows that $K_D(x_1, x_2; \{\lambda, \sigma_1, \sigma_2, x_c\})$ 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_D: by rescaling the input variable appropriately, the change in input scale is removed.

4.4. A sudden change in output scale

Suppose that 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 that we wish to model the function with an isotropic covariance function K of the form (12) 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

$$a(x; x_c) \triangleq \begin{cases} \lambda_1, & x < x_c \\ \lambda_2, & x \geq x_c \end{cases}. \tag{21}$$

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

$$\begin{aligned} &K_E(x_1, x_2; \{\lambda_1^2, \lambda_2^2, \sigma, x_c\}) \\ &\quad \triangleq a(x_1; x_c) a(x_2; x_c) K(x_1, x_2; \{1, \sigma\}) \\ &\quad = \begin{cases} K(x_1, x_2; \{\lambda_1, \sigma\}), & x_1, x_2 < x_c \\ K(x_1, x_2; \{\lambda_2, \sigma\}), & x_1, x_2 \geq x_c \\ K(x_1, x_2; \{(\lambda_1 \lambda_2)^{\frac{1}{2}}, \sigma\}), & \text{otherwise} \end{cases}. \end{aligned} \tag{22}$$

The form of K_E follows from the properties of covariance functions; see [20] for more details.

4.5. A change in observation likelihood

Hitherto, we have taken the observation likelihood $p(\boldsymbol{z} \mid \boldsymbol{y}, \theta, I)$ as being both constant and of the simple independent form represented in (1). We now consider other possible observation models, as motivated by fault detection and removal [19]. A sensor fault essentially implies that the relationship between the underlying, or plant, process y and

the observed values z is temporarily complicated. In situations where a model of the fault is known, the faulty observations need not be discarded; they may still contain valuable information about the plant process. We distinguish *fault removal*, for which the faulty observations are discarded, from *fault recovery*, for which the faulty data are utilized with reference to a model of the fault.

The general observation model we now consider is

$$p(\,z \mid \boldsymbol{y}, \theta, I\,) = \mathrm{N}(z; \mathbf{M}(\boldsymbol{x};\theta)\boldsymbol{y} + \boldsymbol{c}(\boldsymbol{x};\theta), \mathbf{K}_F(\boldsymbol{x},\boldsymbol{x};\theta))\,, \tag{23}$$

which allows us to consider a myriad of possible types of fault modes. Here K_F is a covariance matrix associated with the the fault model, which will likely be different from the covariance over y, K. With this model, we have the posteriors

$$p\big(\boldsymbol{y}_\star \mid z_d, \theta, I\big) = \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big)\,, \mathbf{C}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big)\big)\,, \tag{24}$$

where we have

$$\begin{aligned}\boldsymbol{m}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big) &= \boldsymbol{\mu}(\boldsymbol{x}_\star;\theta) + \mathbf{K}(\boldsymbol{x}_\star, \boldsymbol{x}_d;\theta)\mathbf{M}(\boldsymbol{x}_d;\theta)^{\mathsf{T}} \\ &\quad \times \mathbf{V}_F(\boldsymbol{x}_d, \boldsymbol{x}_d;\theta)^{-1}(z_d - \mathbf{M}(\boldsymbol{x}_d;\theta)\boldsymbol{\mu}(\boldsymbol{x}_d;\theta) \\ &\quad - \boldsymbol{c}(\boldsymbol{x}_d;\theta)) \\ \mathbf{C}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big) &= \mathbf{K}(\boldsymbol{x}_\star, \boldsymbol{x}_\star;\theta) - \mathbf{K}(\boldsymbol{x}_\star, \boldsymbol{x}_d;\theta)\mathbf{M}(\boldsymbol{x}_d;\theta)^{\mathsf{T}} \\ &\quad \times \mathbf{V}_F(\boldsymbol{x}_d, \boldsymbol{x}_d;\theta)^{-1}\mathbf{M}(\boldsymbol{x}_d;\theta)\mathbf{K}(\boldsymbol{x}_d, \boldsymbol{x}_\star;\theta),\end{aligned}$$

and

$$\mathbf{V}_F(\boldsymbol{x}_d, \boldsymbol{x}_d;\theta) \triangleq \mathbf{K}_F(\boldsymbol{x},\boldsymbol{x};\theta) + \mathbf{M}(\boldsymbol{x};\theta)^{\mathsf{T}}\mathbf{K}(\boldsymbol{x},\boldsymbol{x};\theta)\,\mathbf{M}(\boldsymbol{x};\theta).$$

If required, we can also determine the posterior for the fault contributions, defined as $f \triangleq z - y$.

$$\begin{aligned}&p\big(\boldsymbol{f}_\star \mid z_d, \theta, I\big) \\ &\quad = \iint p\big(\boldsymbol{f}_\star \mid z_\star, \boldsymbol{y}_\star, \theta, I\big)\, p\big(z_\star \mid \boldsymbol{y}_\star, \theta, I\big) \\ &\qquad \times p\big(\boldsymbol{y}_\star \mid z_d, \theta, I\big)\,\mathrm{d}\boldsymbol{y}_\star \mathrm{d}z_\star \\ &\quad = \iint \delta\big(\boldsymbol{f}_\star - (z_\star - \boldsymbol{y}_\star)\big) \\ &\qquad \times \mathrm{N}\big(z_\star; \mathbf{M}(\boldsymbol{x}_\star;\theta)\,\boldsymbol{y}_\star + \boldsymbol{c}(\boldsymbol{x};\theta), \mathbf{K}_F(\boldsymbol{x},\boldsymbol{x};\theta)\big)\,\mathrm{d}z_\star \\ &\qquad \times \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big)\,, \mathbf{C}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big)\big)\,\mathrm{d}\boldsymbol{y}_\star \\ &\quad = \mathrm{N}\big(\boldsymbol{f}_\star; \boldsymbol{m}\Big(\mathring{\boldsymbol{f}}_\star \mid z_d, \theta, I\Big)\,, \mathbf{C}\Big(\mathring{\boldsymbol{f}}_\star \mid z_d, \theta, I\Big)\Big),\end{aligned} \tag{25}$$

where we have

$$\begin{aligned}\boldsymbol{m}\Big(\mathring{\boldsymbol{f}}_\star \mid z_d, \theta, I\Big) &= (\mathbf{M}(\boldsymbol{x}_\star;\theta) - \mathbf{E})\boldsymbol{m}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big) + \boldsymbol{c}(\boldsymbol{x}_\star;\theta) \\ \mathbf{C}\Big(\mathring{\boldsymbol{f}}_\star \mid z_d, \theta, I\Big) &= \mathbf{K}_F(\boldsymbol{x}_\star, \boldsymbol{x}_\star;\theta) + (\mathbf{M}(\boldsymbol{x}_\star;\theta) - \mathbf{E}_\star) \\ &\quad \times \mathbf{C}\big(\mathring{\boldsymbol{y}}_\star \mid z_d, \theta, I\big)\,(\mathbf{M}(\boldsymbol{x}_\star;\theta) - \mathbf{E}_\star)^{\mathsf{T}},\end{aligned}$$

where $\mathbf{E}_\star$ is the identity matrix of side length equal to $\boldsymbol{x}_\star$. We now consider some illustrative examples of fault types modelled by this approach.

4.5.1. *Bias*

Perhaps the simplest fault mode is that of *bias*, in which the readings are simply offset from the true values by some constant amount (and then, potentially, further corrupted by additive Gaussian noise). Clearly, knowing the fault model in this case will allow us to extract information from the faulty readings; here we are able to perform fault recovery. In this scenario, $\mathbf{M}(\boldsymbol{x};\theta)$ is the identity matrix, $\mathbf{K}_F(\boldsymbol{x},\boldsymbol{x};\theta)$ is a diagonal matrix whose diagonal elements are identical noise variances (as implicit in (1)) and $\boldsymbol{c}(\boldsymbol{x};\theta)$ is a non-zero constant for x lying in the faulty period, and zero otherwise. The value of the offset and the start and finish times for the fault are additional hyperparameters to be included in θ.

4.5.2. *Stuck value*

Another simple fault model is that of a *stuck value*, in which our faulty readings return a constant value regardless of the actual plant process. We consider the slightly more general model in which those faulty observations may also include a Gaussian noise component on top of the constant value. Here, of course, we can hope only for fault removal; the faulty readings are not at all pertinent to an inference about the underlying variables of interest. This model has, as before, $\mathbf{K}_F(\boldsymbol{x},\boldsymbol{x};\theta)$ equal to a diagonal matrix whose diagonal elements are identical noise variances. $\mathbf{M}(\boldsymbol{x};\theta)$ is another diagonal matrix whose ith diagonal element is equal to zero if x_i is within the faulty region, and is equal to one otherwise. $\mathbf{M}(\boldsymbol{x};\theta)$ hence serves to select only non-faulty readings. $\boldsymbol{c}(\boldsymbol{x};\theta)$, then, is equal to a constant value (the stuck value) if x_i is within the faulty region, and is equal to zero otherwise. Here, as for the biased case, we have additional hyperparameters corresponding to the stuck value and the start and finish times of the fault.

4.5.3. *Drift*

The final fault we consider is that of drift. Here our sensor readings undergo a smooth excursion from the plant process; that is, they gradually 'drift' away from the real values, before eventually returning back to normality. Unsurprisingly, here $\mathbf{K}_F(\boldsymbol{x},\boldsymbol{x};\theta)$ is a drift covariance K_C as defined in (18), with the addition of noise variance terms to its diagonal as required. Otherwise, $\mathbf{M}(\boldsymbol{x};\theta)$ is the appropriate identity matrix and $\boldsymbol{c}(\boldsymbol{x};\theta)$ is a zero vector. With knowledge of this model, fault recovery is certainly possible. The model requires additional parameters that define the relevant covariance K used in (18), as well as the fault start and finish times.

4.6. 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 (10), 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 (11) and (12).

As such, we are able to calculate the posterior distributions of any unknown quantity, such as the putative location of a changepoint, x_c, or the probability that a fault of a particular type might have occurred. In some applications, it may be necessary to make a hard decision, that is, to commit to a changepoint having occurred at a given point in time. This would be necessary, for example, if a system had correctional or responsive actions that it could take when a changepoint occurs. Fortunately, we can address the temporal segmentation problem using simple Bayesian decision theory. Given our observations $(\boldsymbol{x}_d, \boldsymbol{z}_d)$, we can determine the probability that there was a changepoint at x_c, $P(\text{Changepoint}(x_c) \mid \boldsymbol{z}_d, I)$, using (11). Now after specifying the costs of false positive and false negative changepoint reports as c_{I} and c_{II}, respectively (and taking the cost of true positive and true negative reports as zero), we can take the action that minimizes the expected loss. If $(1 - P(\text{Changepoint}(x_c) \mid \boldsymbol{z}_d, I))c_{\text{I}} < P(\text{Changepoint}(x_c) \mid \boldsymbol{z}_d, I)\, c_{\text{II}}$, we specify a changepoint at time x_c; otherwise, we do not. Continuing in this manner, we can segment the entire data stream.

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. If no changepoint is present in the window, the posterior distribution for its location will typically be concentrated at its trailing edge. A changepoint at such a location will have no influence on the predictions; the model is hence able to effectively manage the absence of changepoints.

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 and output scales could be readily modelled by

$$K_G(x_1, x_2; \{\lambda_1, \lambda_2, \sigma_1, \sigma_2, x_c\}) \triangleq a(x_1; x_c)a(x_2; x_c)K(u(x_1; x_c), u(x_2; x_c); \{1, 1\}), \quad (26)$$

where u is as defined in (20) and a is as defined in (21).

For such models, we may be required to decide which type of changepoint to report. Exactly as per our discussion on decisions above, this would require the specification of a loss function, that would, for example, stipulate the loss associated with reporting a change in input scale when there was actually a change in output scale. Given that, we again simply make the report that minimizes our expected loss.

Note 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 its parameters as hyperparameters of the model, and place appropriate hyperparameter samples corresponding to, for example, the constant mean value 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.

5. EXAMPLE

As an expository example, we consider a function that undergoes a sudden change in both input and output scales. The function $y(x)$ is displayed in Fig. 3; 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

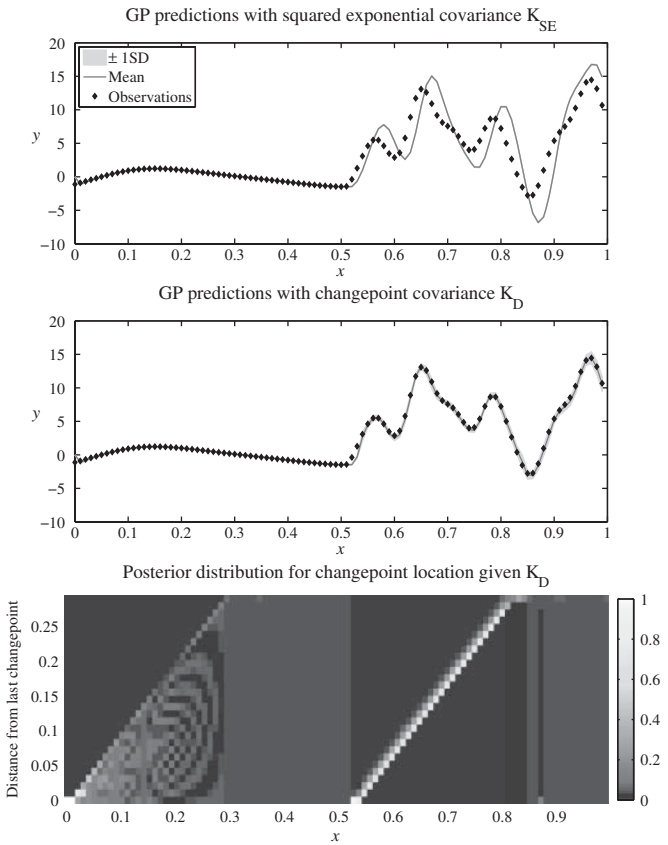

FIGURE 3. Prediction over a function that undergoes a change in both input and output scales using covariance K_G.

of performing one-step lookahead prediction on $y(x)$ using GP models with a moving window of size 25.

The uppermost plot in Fig. 3 shows the performance of a standard GP prediction model with the squared exponential covariance K_{SE} (14), using hyperparameters $\{\lambda, \sigma\}$ selected by maximum-likelihood-II 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 predictive variance) that are nonetheless very inaccurate.

The central plot shows the performance of a GP prediction model using the changepoint covariance function K_G (26). The predictions were calculated via BQ hyperparameter marginalization using (10); 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 data set, the minimum water levels of the Nile river during the period AD 622–1284 [23]. Several authors have found evidence supporting a change in input scale for this data around the year AD 722 [8]. The conjectured reason for this changepoint is the construction in AD 715 of a 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 data set using the input scale changepoint covariance K_D (19), and a moving window of size 150. Eleven samples each were used for the hyperparameters σ_1 and σ_2, the input scales before and after a putative changepoint, respectively, and 150 samples were used for the location of the changepoint x_c.

The results can be seen in Fig. 4. The upper plot shows our predictions for the data set, 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.

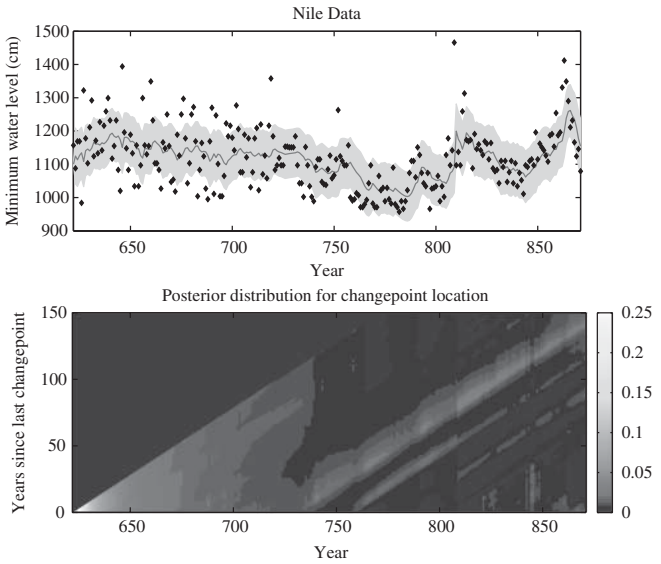

FIGURE 4. Prediction for the Nile data set using input scale changepoint covariance K_D, and the corresponding posterior distribution for time since changepoint.

6.2. Well-log data

Also commonly considered in the context of changepoint detection is the *well-log* data set, comprising 4050 measurements of nuclear magnetic response made during the drilling of a well [24]. The changes here correspond to the transitions between different strata of rock.

We performed prediction on this data set using a simple diagonal covariance that assumed that all measurements were independent and identically distributed (IID). The noise variance for this covariance (alternatively put, its output scale) was determined by maximum likelihood; it was assumed known *a priori*. We then took a mean function that was constant for each rock stratum; that is, the mean undergoes changes at changepoints (and only at changepoints). Given the length of the data set, and that regions of data before and after a changepoint are independent, we performed predictions for a point by considering a window of data centred on that point. Essentially, we performed sequential prediction for predictants midway through the window. In each window (comprising 50 observations), we allowed for a single changepoint. Hence, our model was required to marginalize over three hyperparameters, the mean before the changepoint, the mean after the changepoint and the location of that changepoint. For these hyperparameters, we took 13, 13 and 40 samples, respectively.

We compared our results against those produced by a variational Bayesian hidden Markov model with a mixture of Gaussian's emission probability [25, 26]. This model gave a log marginal likelihood of $\log p(z_d|I) \simeq -1.51 \times 10^5$, whereas our GP model gave $\log p(z_d|I) \simeq -1.02 \times 10^4$. The resulting predictions for both methods are depicted in Fig. 5. According to our metric, our GP model's performance was an order of

magnitude better than this alternative method, largely due to the predictions made in the regions just prior to $x = 1600$ and just after $x = 2400$.

6.3. 1972–1975 Dow–Jones industrial average

A final canonical changepoint data set is the series of daily returns of the Dow-Jones industrial average between the 3 July 1972 and the 30 June 1975 [6]. This period included a number of newsworthy events that had significant macroeconomic influence, as reflected in the Dow–Jones returns.

We performed sequential prediction on this data using a GP with a diagonal covariance that assumed all measurements were IID. However, the variance of these observations was assumed to undergo changes, and as such we used a covariance K_D that incorporated such changes in the output scale. The window used was 350 observations long, and was assumed to contain no more than a single changepoint. As such, we had three hyperparameters to marginalize: the variance before the changepoint, the variance after the changepoint and, finally, the location of that changepoint. For these hyperparameters, we took 50, 17 and 17 samples, respectively.

Our results are plotted in Fig. 6. Our model clearly identifies the important changepoints that likely correspond to the commencement of the OPEC embargo on the 19 October 1973, and the resignation of Richard Nixon as President of the USA

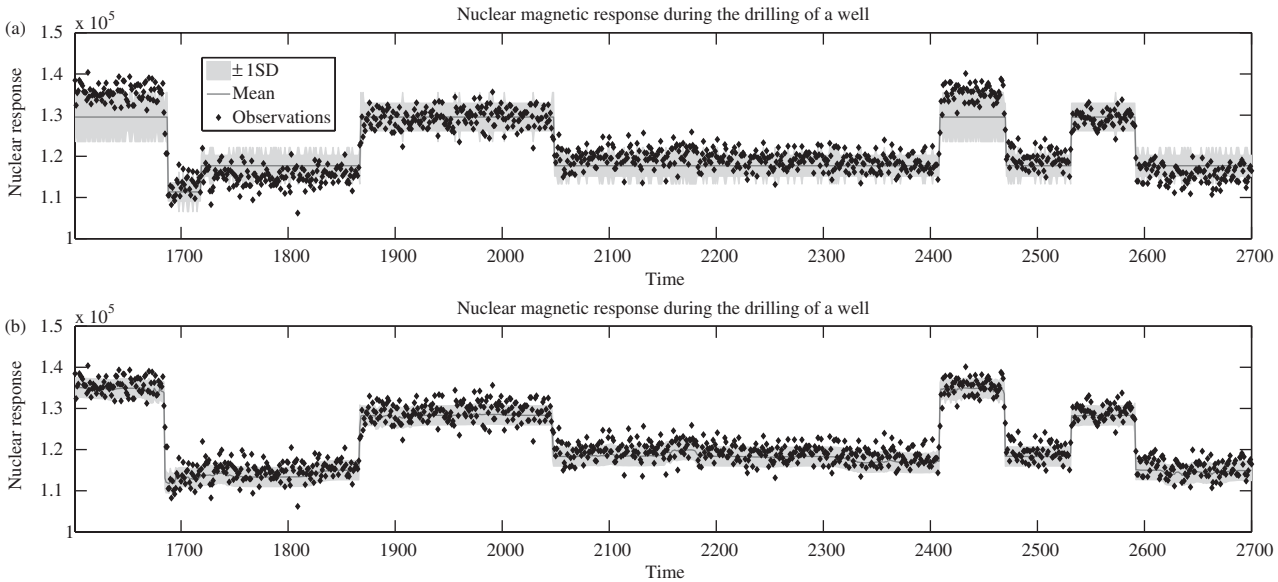

FIGURE 5. Retrospective predictions for the well-log data using (a) hidden Markov model and (b) a GP with drastic change covariance function K_A.

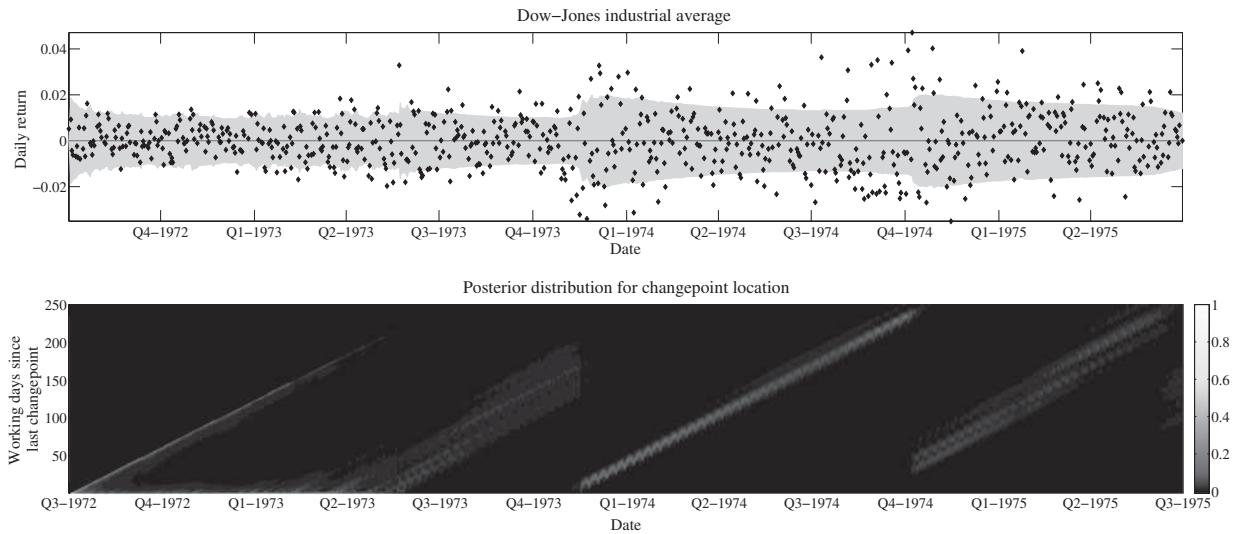

FIGURE 6. Online predictions and posterior for the location of changepoint for the Dow-Jones industrial average data using covariance K_D.

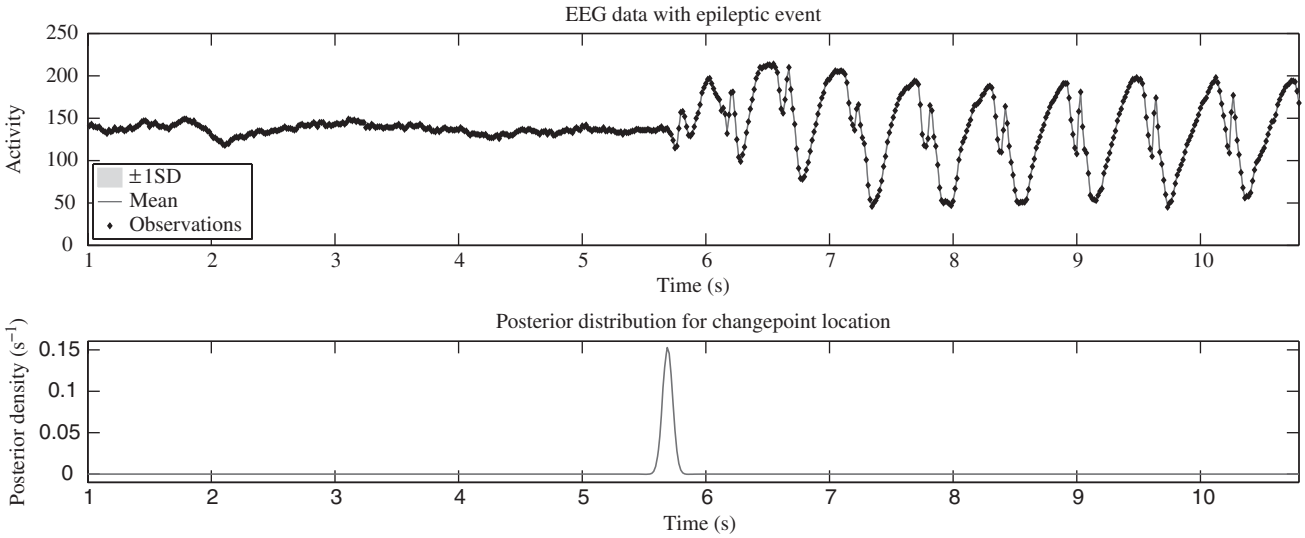

FIGURE 7. Retrospective predictions and posterior for the location of changepoint for the EEG data with epileptic event. The covariance K_B was employed within our GP framework.

on the 9 August 1974. A weaker changepoint is identified early in 1973, which [6] speculate is due to the beginning of the Watergate scandal.

6.4. Electroencephalography data with epileptic event

We now consider electroencephalography (EEG) data from an epileptic subject [27]. Prediction here is performed with the aim of ultimately building models for EEG activity strong enough to forecast seizure events [28]. The particular data set plotted in Fig. 7 depicts a single EEG channel recorded at 64 Hz with 12-bit resolution. It depicts a single epileptic event of the classic 'spike and wave' type.

We used the covariance K_B (17) to model our data, accommodating the smooth transition of the data between drastically different regimes. We took K_1 as a simple squared exponential (14) and K_2 as a periodic covariance (15) multiplied by another squared exponential. K_2 is intended to model EEG data during the course of seizure, K_1, data from other regions. We assume that we have sufficient exemplars of EEG data unaffected by seizure to set the hyperparameters for K_1 using maximum likelihood. We further assumed that the input scale of the non-periodic squared exponential within K_2 was identical to that for K_1, representing a constant long-term smoothness for both seizure and non-seizure periods. The hyperparameters we were required to marginalize, then, were the period σ, amplitude λ and smoothness ω of (15) for K_2, along with the location of the changepoint and its type (either periodic to non-periodic or non-periodic to periodic). For these hyperparameters, we took, respectively, 7, 7, 5, 50 and 2 samples.

This model was used to perform effective retrospective prediction over the data set, as depicted in Fig. 7. As can be seen, our posterior distribution for the location of the changepoint correctly locates the onset of seizure.

6.5. Stuck sensor

To illustrate our approach to sensor fault detection, we also tested on a network of weather sensors located on the south coast of England.[2] We considered the readings from the Sotonmet sensor, which makes measurements of a number of environmental variables (including wind speed and direction, air temperature, sea temperature and tide height) and makes up-to-date sensor measurements available through separate web pages (see http://www.sotonmet.co.uk). This sensor is subject to network outages and other faults that suggest the use of the models described in Section 4.5.

In particular, we performed on-line prediction over tide height data in which readings from the sensor became stuck at an incorrect value. As such, we used the change in the observation model taken from Section 4.5.2. The covariance for the underlying plant process was taken to be the sum of a periodic and a non-periodic component, as described in [21], the hyperparameters for which can be determined off-line. As such, we need to marginalize only the hyperparameter corresponding to the location of a changepoint in the window, and over the type of that change point (i.e. either not-stuck to stuck or stuck to not-stuck). Clearly, our belief about the stuck value can be heuristically determined for any appropriate region — it is a delta distribution at the constant observed value. We employed a window size of 350 data points, and, correspondingly, 350

[2]The network is maintained by the Bramblemet/Chimet Support Group and funded by organizations including the Royal National Lifeboat Institution, Solent Cruising and Racing Association and Associated British Ports.

samples for the location of the changepoint. Results are plotted in Fig. 8. Our model correctly identified the beginning and end of the fault. Then by performing fault removal via (24), the model is able to perform effective prediction for the plant (tide) process throughout the faulty region.

6.6. EEG data with saccade event

To illustrate our approach to sensor fault recovery, we also tested on a Brain-Computer Interface (BCI) application. BCI can be used for assisting sensory-motor functions as well as monitoring sleep patterns. EEG is a highly effective non-invasive interface. However, the EEG signal can often be corrupted by electro-oculogram (EOG) artefacts that may be the result of a saccade; it is necessary to remove the artefact from the EEG signal. This problem was treated as a blind source separation problem in [29] and an ICA solution was proposed which identified the separate artefact-free EEG signal (which we refer to as EEG*) and the EOG signal. Figure 9 shows typical EOG activity during a saccade. In BCI applications, however, a measured EOG signal is rarely available and we must rely on the artifact removal algorithms to offer an accurate assessment of the pure EEG* signal.

We demonstrate an alternative approach to EOG artefact removal that we first proposed in [19]. Our approach allows the user to encode any available information about the shape of the component signals including signal smoothness, signal continuity at change points and even the shape of the signal if sufficient training data is available. In our approach, both the EEG* and EOG signals are modelled using GPs and these signals are determined from the EEG signal data using the fault recovery approach outlined in Section 4.5. Although the application of GPs to artefact detection in EEG signals is not new [28], as far as we can see, the use of GPs to actively remove the artefact and thus recover the underlying pure EEG signal is novel.

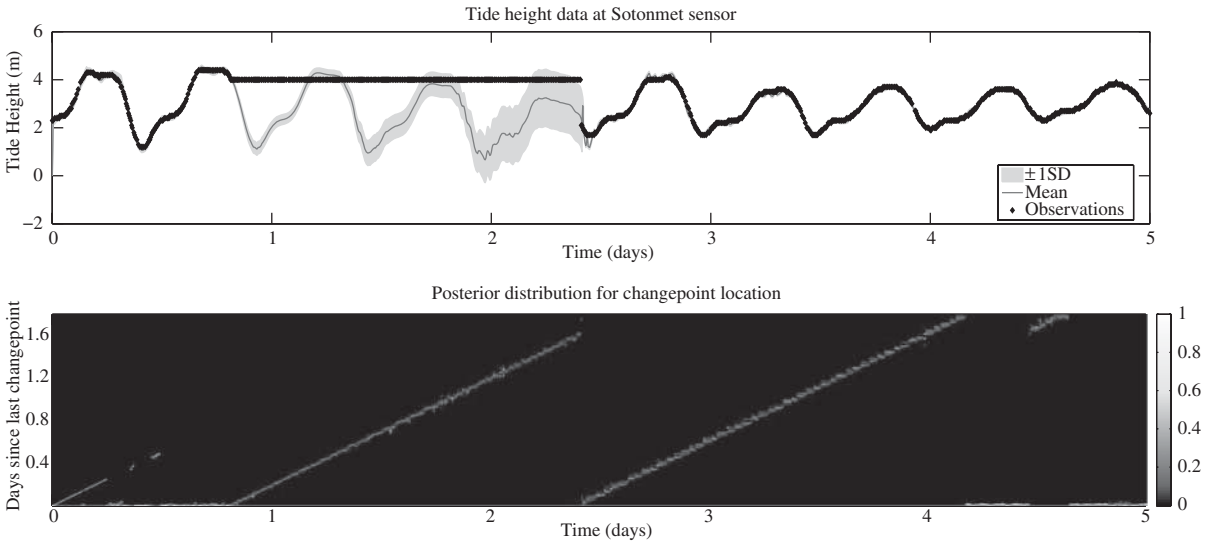

FIGURE 8. Online predictions and posterior for the location of changepoint for the tide height data. The fault was modelled as a change in observation likelihood of the form described in Section 4.5.2.

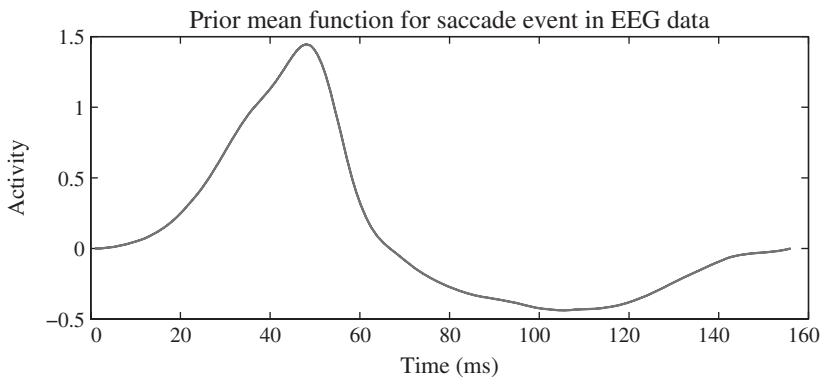

FIGURE 9. EOG activity during a saccade event.

The EEG* signal is modelled as a smooth function generated from a squared exponential covariance function. The EOG signal is a function that undergoes a temporary excursion from an otherwise constant value of zero and, as such, is modelled using the 'drift' model, (18). We shall, however, consider two variations of the drift model when modelling the EOG artefact. These variations differ only in the prior mean that is assigned to the EOG artefact model. The first variation assumes that no further information about the shape of the EOG signal is known and, in this case, the EOG artefact prior mean is zero throughout. For the second variation of the drift model, a prior mean is learnt from samples of EOG signals, giving the shape depicted in Fig. 9. In this case, the EOG covariance function models the residual between the prior EOG mean and the current signal.

The presence of abundant uncorrupted EEG signal data allowed the length and height scale hyperparameters for the EEG* model to be learnt using maximum likelihood. We modelled the dynamics of the EOG excursion itself using a squared exponential covariance function, and assumed that its input scale was the same as for the EEG data. As such, we were required to marginalize three hyperparameters: the output scale λ of the EOG covariance, and the artefact start time and duration. For the zero mean fault model we took 13, 13 and 150 samples, respectively, for those hyperparameters. For the non-zero mean model we took 5, 7 and 75 samples, respectively. The non-zero mean model also requires a vertical scaling factor for the prior mean shape (Fig. 9) and, for this hyperparameter, we took nine samples.

For the artefact start time hyperparameter, we took a uniform prior over the extent of the dataset. As usual, if no artefact was detected, the posterior mass for the start time would be concentrated at the end at the data set. We cannot be very certain *a priori* as to the duration of a saccade, which will be dependent upon many factors (notably, the size of the saccade). However, a reasonable prior [30] might place upon the logarithm of the saccade duration a Gaussian with a mean of log(110 ms) and a standard deviation of 0.6 (meaning that saccade durations of 60 and 200 ms are both a single SD from the mean). This was the prior taken for the artefact duration hyperparameter.

Figures 10 and 11 show the result of performing retrospective prediction over our EEG data. Figure 10 shows the 1 standard error confidence interval for the artefact-free EEG* signal and the EOG artefact obtained using our algorithm with a zero prior mean EOG model. The figure also shows the retrospective posterior distribution over the artefact start time. Although our approach is able to determine when an artefact occurs, its start

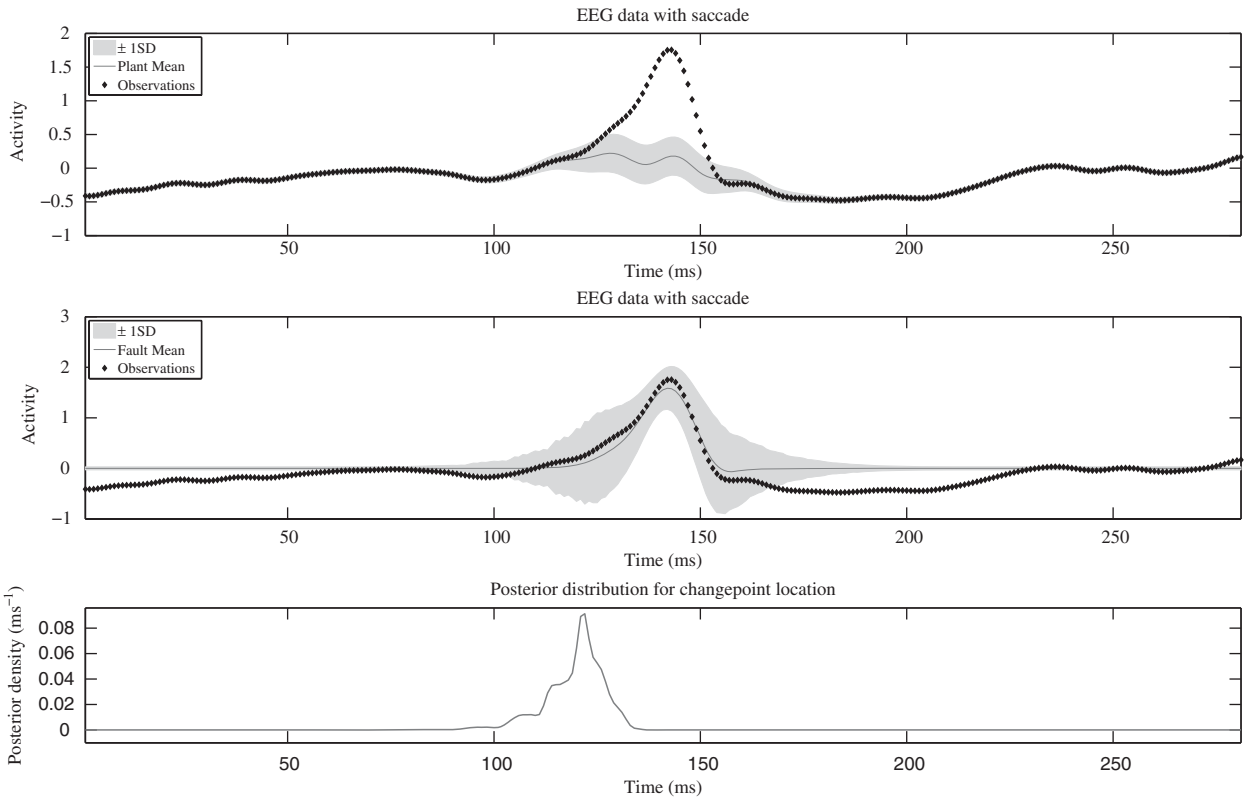

FIGURE 10. Retrospective predictions and posterior for the location of changepoint for the EEG data with saccade. Predictions are made both for the plant process (the underlying EEG signal) using (24), as well as for the fault contribution due to saccade, using (25). The GP assumes a zero prior mean during the saccade.

time is hard to determine as, at the artefact onset, the EEG signal length scale is similar to the pure EEG* signal. However, the approach successfully removes the EOG artefact from the EEG signal. We can also use (10) to produce the full posterior for the EEG signal over the saccade event, as plotted in Fig. 12a. Note that we can distinguish two models: the model that simply follows the EEG signal; and the model that assumes that a saccade artefact may have occurred. The former is characterized by a tight distribution around the observations, the latter being much more uncertain due to its assumption of a fault. Note that the first model gradually loses probability mass to the second until the first becomes completely implausible.

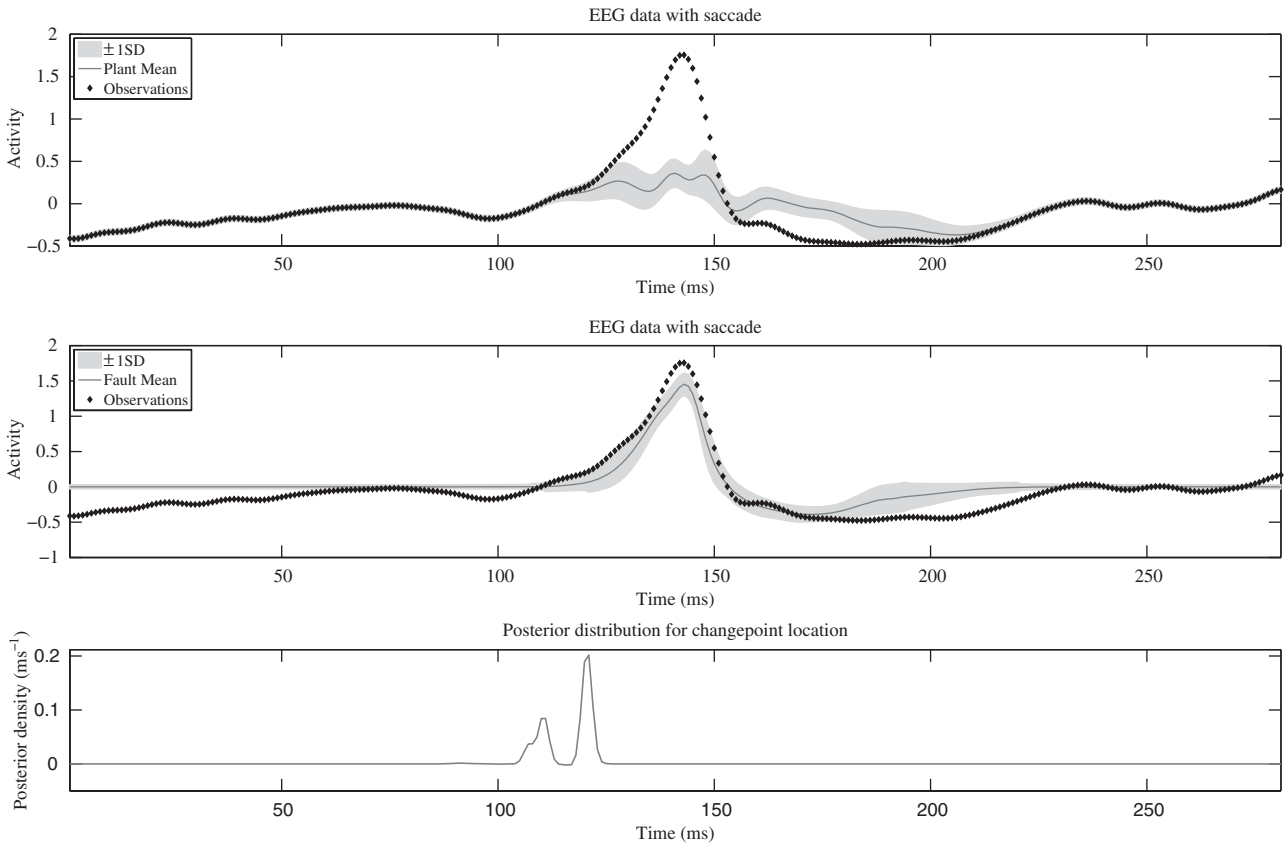

FIGURE 11. Retrospective predictions and posterior for the location of changepoint for the EEG data with saccade. Predictions are made both for the plant process (the underlying EEG signal) using (24), as well as for the fault contribution due to saccade, using (25). The GP assumes a prior mean during the saccade of the common form for EOG activity during such an event.

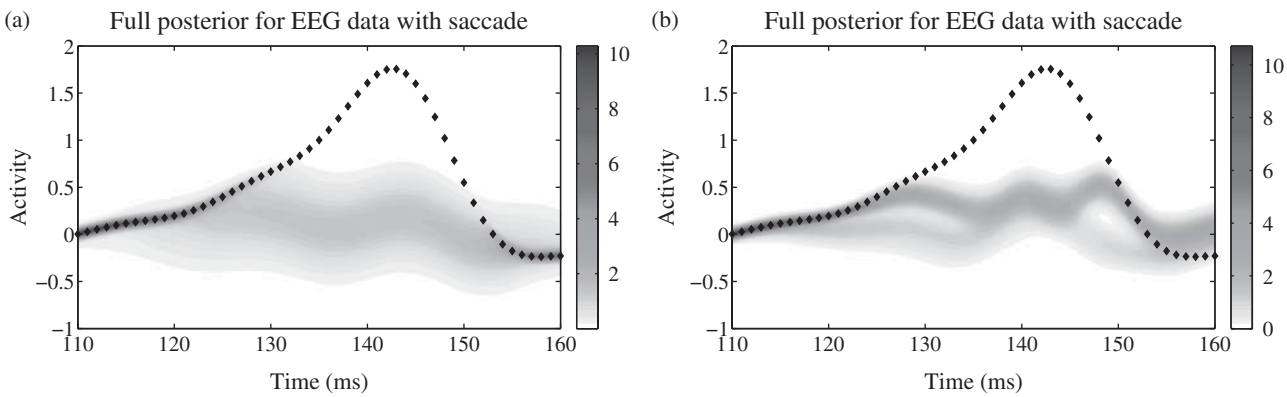

FIGURE 12. Full retrospective posteriors for the EEG data with saccade plant process, using (24). A GP was taken that assumes (a) a zero prior mean and (b) a prior mean during the saccade of the common form for EOG activity during such an event.

Figure 11 shows the recovered signals obtained using the non-zero mean EOG artefact model. In this case, our approach more accurately identifies the start and finish times of the artefact and also accurately separates the pure EEG* and EOG signals. It is interesting to note that this approach results in a bimodal distribution over the artefact start time. The most likely start times are identified by our algorithm to occur at the kinks in the data at t = 115 ms and t = 120 ms. This results in a bimodal estimate of the EEG* and EOG signals. These bimodal signal estimates and the distribution over the artefact start times are shown in Fig. 12b.

7. CONCLUSION

We introduce a new sequential algorithm for performing Bayesian time-series prediction in the presence of changepoints or faults. After developing a variety of suitable covariance functions, 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 data sets demonstrate the efficacy of our algorithm.

ACKNOWLEDGEMENTS

We would like to thank Dr. Hyoung-joo Lee for providing the results of the hidden Markov model over the well-log data.

FUNDING

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

[1] Basseville, M. and Nikiforov, I. (1993) *Detection of Abrupt Changes: Theory and Application*. Prentice Hall.

[2] Brodsky, B. and Darkhovsky, B. (1993) *Nonparametric Methods in Change-Point Problems*. Springer.

[3] Csorgo, M. and Horvath, L. (1997) *Limit Theorems in Change-Point Analysis*. John Wiley & Sons.

[4] Chen, J. and Gupta, A. (2000) *Parametric Statistical Change Point Analysis*. Birkhäuser Verlag.

[5] Chernoff, H. and Zacks, S. (1964) Estimating the current mean of a normally distributed variable which is subject to changes in time. *Ann. Math. Stat.*, **35**, 999–1028.

[6] Adams, R.P. and MacKay, D.J. (2007) Bayesian Online Changepoint Detection. Technical Report, University of Cambridge, Cambridge, UK. arXiv:0710.3742v1 [stat.ML].

[7] Carlin, B.P., Gelfand, A.E. and Smith, A.F.M. (1992) Hierarchical Bayesian analysis of changepoint problems. *Appl. Stat.*, **41**, 389–405.

[8] Ray, B. and Tsay, R. (2002) Bayesian methods for change-point detection in long-range dependent processes. *J. Time Ser. Anal.*, **23**, 687–705.

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

[10] Horváth, L. and Kokoszka, P. (1997) The effect of long-range dependence on change-point estimators. *J. Stat. Plan. Inference*, **64**, 57–81.

[11] Fearnhead, P. and Liu, Z. (2007) On-line inference for multiple changepoint problems. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)*, **69**, 589–605.

[12] Venkatasubramanian, V., Rengaswamy, R., Yin, K. and Kavuri, S. (2003) A review of process fault detection and diagonsis. part 1: Quantitative model-based methods. *Comput. Chem. Eng.*, **27**, 293–311.

[13] Willsky, A. (1976) A survey of design methods for failure detection in dynamic systems. *Automatica*, **12**, 601–611.

[14] Basseville, M. (1988) Detecting changes in signals and systems—a survey. *Automatica*, **24**, 309–326.

[15] Kobayashi, T. and Simon, D. (2003) Application of a Bank of Kalman Filters for Aircraft Engine Fault Diagnosis. *Proc. ASME Turbo Expo 2003, Power for Land, Sea, and Air,* Atlanta, Georgia, USA, June 16–19, 2003.

[16] Aggarwal, V., Nagarajan, K. and Slatton, K. (2004) Estimating Failure Modes Using a Multiple-Model Kalman Filter. Technical Report no. Rep_2004-03-001, ASPL.

[17] Reece, S., Claxton, C., Nicholson, D. and Roberts, S.J. (2009a) Multi-sensor Fault Recovery in the Presence of Known and Unknown Fault Types. *Proc. 12th Int. Conf. Information Fusion (FUSION 2009)*, Seattle, USA.

[18] Garnett, R., Osborne, M.A. and Roberts, S. (2009) Sequential Bayesian Prediction in the Presence of Changepoints. *Proc. 26th Annual Int. Conf. Machine Learning*, Montreal, Canada.

[19] Reece, S., Garnett, R., Osborne, M.A. and Roberts, S.J. (2009b) Anomaly Detection and Removal Using Non-stationary Gaussian Processes. Technical Report, University of Oxford, Oxford, UK.

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

[21] Osborne, M.A., Rogers, A., Ramchurn, S., Roberts, S.J. and Jennings, N.R. (2008) Towards Real-Time Information Processing of Sensor Network Data Using Computationally Efficient Multi-output Gaussian Processes. *Int. Conf. Information Processing in Sensor Networks 2008*, St. Louis, MO, USA, pp. 109–120.

[22] O'Hagan, A. (1991) Bayes-hermite quadrature. *J. Stat. Plan. Inference*, **29**, 245–260.

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

[24] Ruanaidh, J., Fitzgerald, W. and Pope, K. (1994) Recursive Bayesian Location of a Discontinuity in Time Series. *Proc.*

Acoustics, Speech, and Signal Processing, 1994 on IEEE Int. Conf., Vol. 04, pp. 513–516.

[25] Ji, S., Krishnapuram, B. and Carin, L. (2006) Variational Bayes for continuous hidden Markov models and its application to active learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 522–532.

[26] Lee, H. (2009) Variational Bayesian Hidden Markov Models with Mixtures of Gaussian Emission. Technical Report, University of Oxford, Oxford, UK. http://www.robots.ox.ac.uk/∼mosb/Lee2009.pdf.

[27] Roberts, S.J. (2000) Extreme Value Statistics for Novelty Detection in Biomedical Data Processing. *Science, IEE Proc. Sci. Meas. Technol.*, pp. 363–367.

[28] Faul, S., Gregorcic, G., Boylan, G., Marnane, W., Lightbody, G. and Connolly, S. (2007) Gaussian process modeling of EEG for the detection of neonatal seizures. *IEEE Trans. Biomed. Eng.*, **54**, 2151–2162.

[29] Roberts, S.J., Everson, R., Rezek, I., Anderer, P. and Schlögl, A. (1999) Tracking ICA for Eye-Movement Artefact Removal. *Proc. EMBEC'99*, Vienna.

[30] Jürgens, R., Becker, W. and Kornhuber, H. (1981) Natural and drug-induced variations of velocity and duration of human saccadic eye movements: evidence for a control of the neural pulse generator by local feedback. *Biol. Cyber.*, **39**, 87–96.