Active Data Selection for Sensor Networks with Faults and Changepoints

Michael A. Osborne, Roman Garnett, Stephen J. Roberts
Department of Engineering Science
University of Oxford
Oxford, UK, OX1 3PJ
Email: {mosb, rgarnett, sjrob}@robots.ox.ac.uk

***Abstract*—We describe a Bayesian formalism for the intelligent selection of observations from sensor networks that may intermittently undergo faults or changepoints. Such active data selection is performed with the goal of taking as few observations as necessary in order to maintain a reasonable level of uncertainty about the variables of interest. The presence of faults/changepoints is not always obvious and therefore our algorithm must first detect their occurrence. Having done so, our selection of observations must be appropriately altered. Faults corrupt our observations, reducing their impact; changepoints (abrupt changes in the characteristics of data) may require the transition to an entirely different sampling schedule. Our solution is to employ a Gaussian process formalism that allows for sequential time-series prediction about variables of interest along with a decision theoretic approach to the problem of selecting observations.**

***Keywords*-active data selection, sensor selection, sensor networks, Gaussian processes, time-series prediction, changepoint detection, fault detection, Bayesian methods**

I. Introduction

Active data selection is the task of selecting only the most informative observations of a system under study. Specifically, we aim to select observations that render us as certain as possible about some variable(s) of interest. In this paper, we consider networks of sensors that are all capable of making observations of various correlated variables, such as the air temperatures at the different sensor locations. Our goals in this scenario are twofold: we want to select observations that minimise our uncertainty about those variables, while also minimising the required number of those observations. Note that in practice, taking observations is usually associated with some cost, such as the battery energy required to power a sensor or the computational cost associated with additional data. An observation might also be associated with an increased risk of detection, if our system's goal is to perform covert surveillance.

Active data selection has been the topic of much previous research [1], [2]. Unlike that work, we consider the problem of active data selection for on-line time-series prediction. This problem benefits from a clear definition of the points of our interest—at any time t, we make predictions about the values of a field at every sensor location for a single specified time $t + \epsilon$. Here ϵ is the *lookahead*, which for simple tracking is zero. Our first goal can be restated as the minimisation of the sum of the standard deviations of those predictions.

Making those predictions, however, is complicated by *changepoints* in our variables. Changepoints are abrupt changes to the properties of the measured data. 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. 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. Typically we receive no definite notification of the occurrence of such changepoints; our algorithm must be sufficiently flexible to recognise the characteristic changes in the observations received and treat them appropriately.

The problem of detecting and locating changepoints has been widely studied. A large number of methods have been proposed for this problem; see [3]–[8] and the references therein for more information. However, very little work [9] has been directed towards the problem of on-line prediction where such changepoints might exist. Some of these approaches [10]–[12] employ Kalman filters to perform fault recognition, and, subsequently, prediction. We build upon recent work that has taken Gaussian processes, a generalisation of the Kalman filter approach, for prediction in the presence of changepoints [13], [14].

Active data selection using Gaussian processes has received previous attention [15]. We aim to extend that work to manage more problematic datasets containing changepoints and faults. This presents several novel challenges. If a changepoint is detected, this will typically require the transition to an appropriately different sampling schedule. The detection of the changepoint will, in itself, present additional difficulties for the active data selection procedure. For example, if we suspect that a fault has either just occurred or just ended, we might want observations to be selected in order to confirm that suspicion.

We set our problem within a fully Bayesian framework. The locations, types and characteristics of changepoints become hyperparameters of our model, which can be marginalised out using numerical integration techniques to give the posterior distribution for the variables of our interest. Additionally, posterior distributions for those hyperparameters can be produced as desired. Given our probabilistic model, we can then specify the active data selection task as a problem of decision theory, using a loss

function designed to match the twin goals outlined above. This gives us the Bayesian solution to the selection of data from sources that may contain changepoints.

II. Gaussian process prediction in the presence of changepoints

Gaussian processes (GPs) offer a powerful method to perform Bayesian inference about functions [16]. A GP is defined as a distribution over the functions $X \to \mathbb{R}$ such that the distribution over the possible function values on any finite subset of X is multivariate Gaussian. For a function $y(x)$, the prior distribution over its values $\boldsymbol{y}$ on a subset $\boldsymbol{x} \subset X$ are completely specified by a mean vector $\boldsymbol{\mu}(\boldsymbol{x})$ and covariance matrix $\mathbf{K}(\boldsymbol{x}, \boldsymbol{x})$,

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

Here I, the *context*, includes prior knowledge of both the mean and covariance functions, which generate $\boldsymbol{\mu}$ and $\mathbf{K}$ respectively. We will incorporate knowledge of relevant functional inputs, such as x, into I for notational convenience. The prior mean function is chosen as appropriate for the problem at hand (often a constant), and the covariance function is chosen to reflect any prior knowledge about the structure of the function of interest, for example periodicity.

Popular examples are the squared exponential covariance function, given by

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

or the the Matérn class of covariance functions, parameterised by a 'roughness' ν, which for the example of $\nu = \frac{3}{2}$ can be written

$$\begin{aligned} &K^{(\mathrm{Mtn})}(x_1, x_2; \lambda, \sigma, \nu = 3/2) \triangleq \\ &\quad \lambda^2 \left(1 + \sqrt{3}\,\frac{|x_1 - x_2|}{\sigma}\right) \exp\left(-\sqrt{3}\,\frac{|x_1 - x_2|}{\sigma}\right). \end{aligned} \tag{2}$$

The parameters λ and σ represent respectively the characteristic *output* and *input scales* of the process. λ and σ are examples of the set of hyperparameters, collectively denoted as θ, that are required to specify our covariance and mean functions. Other covariance functions can be constructed for a wide variety of problems [16]. For this reason, GPs are ideally suited for time-series prediction problems with complex behavior.

In the context of this paper, we will take y to be a potentially dependent dynamic process, such that X contains a time dimension. Note that our approach considers functions of continuous time; we have no need to discretise our observations into time steps. In particular, when considering sensor networks, our inputs are of the form $x = [l, t]$, for *sensor label* l (if we have L different sensors, $l = 1, \ldots, L$) and time t. We construct a covariance function over this input space by simply taking the product of a covariance term over label alone and a term over time alone, as

$$K([l_1, t_1], [l_2, t_2]) \triangleq K^{(l)}(l_1, l_2)\, K^{(t)}(t_1, t_2)\,, \tag{3}$$

The covariance over label can be found by considering, for example, a function of the spatial separation between sensors, or, if the number of sensors is not too large, can be arbitrarily parameterised [15]. Note that l can also be used to label the different sensing modes of a single sensor; different l need not necessarily correspond to different spatial locations.

The changepoints we wish to consider will exist only within the term over time. As such, we require covariance models only for changepoints in a one-dimensional input t. Such covariance functions exist to model changepoints and faults of many different types [17], some examples of which are displayed in Figure 1. The extension to changepoints over higher-dimensional spaces is straightforward given an appropriate parameterisation of the boundary between regions of different characteristics. Changepoint covariances are also specified by hyperparameters, such as the location and type of each changepoint, which will be included within θ.

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

$$V(t_1, t_2; \theta) \triangleq K(t_1, t_2; \theta) + \eta^2\, \delta(t_1 - t_2)\,, \tag{4}$$

where $\delta(\cdot)$ is the Kronecker delta function. Of course, in the noiseless case, $z = y$ and $V(t_1, t_2; \theta) = K(t_1, t_2; \theta)$.

Where we have faults, (4) must be appropriately modified [17]. A sensor fault essentially implies that the relationship between the underlying, or plant, process y and the observed values z is temporarily altered. As such, $p(\,\boldsymbol{z} \mid \boldsymbol{y}, \theta, I\,)$ will express non-stationary, dependent noise contributions. In order to describe such faults, we include additional hyperparameters into θ specifying their time of occurrence, duration and type. In situations where a model of the fault is known, the faulty observations need not be discarded. Our principled probabilistic approach will allow us to extract whatever information faulty observations may contain that is pertinent to inference about the plant process.

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

$$\begin{aligned} &p(\boldsymbol{y}_\star \mid \boldsymbol{z}_d,\, \theta,\, I) \\ &\qquad = \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta,\, I)\,, \mathbf{C}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta,\, I)\big)\,, \end{aligned} \tag{5}$$

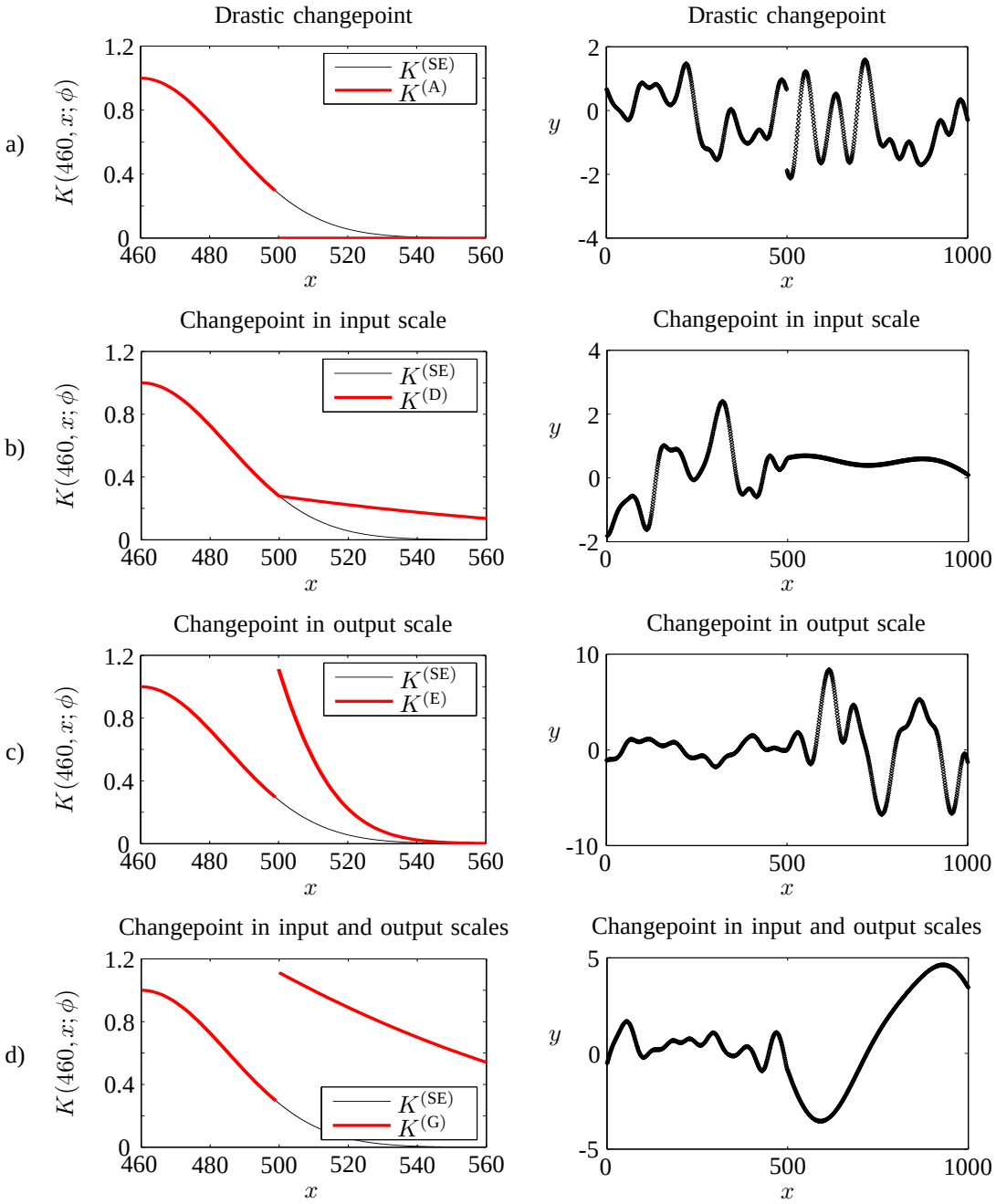

Figure 1: The squared exponential covariance (1) compared with example covariance functions for the modelling of data with changepoints [17], and associated example data that they might be appropriate for.

where we have[1]

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

We use the formulation of a GP given by [15] to perform sequential prediction using an adaptive moving window. After each new observation, we use rank-one updates to the covariance matrix to efficiently update our predictions in light of the new information received. We efficiently remove the trailing edge of the window using a similar rank-one "downdate." The computational savings made by these choices mean our algorithm can be feasibly run online.

Of course, we can rarely be certain about θ *a priori*. These hyperparameters must therefore be marginalised. Although the required integrals are non-analytic, we can efficiently approximate them by use of Bayesian Monte Carlo [18] techniques. This entails evaluating our predictions for a range of hyperparameter samples $\{\theta_i : i \in S\}$, with a different mean $\boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)$ and covariance $\mathbf{C}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)$ for each, which are then combined in

[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.

a weighted mixture

$$
\begin{aligned}
&p(\,\boldsymbol{y}_\star \mid \boldsymbol{z}_d,\, I\,) \\
&= \frac{\int p(\,\boldsymbol{y}_\star \mid \boldsymbol{z}_d, \theta, I\,)\, p(\,\boldsymbol{z}_d \mid \theta, I\,)\, p(\,\theta \mid I\,)\,\mathrm{d}\theta}{\int p(\,\boldsymbol{z}_d \mid \theta, I\,)\, p(\,\theta \mid I\,)\,\mathrm{d}\theta} \\
&\simeq \sum_{i \in S} \rho_i\, \mathrm{N}\big(\boldsymbol{y}_\star; \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)\,, \mathbf{C}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)\big)\,,
\end{aligned}
\tag{6}
$$

with weights $\boldsymbol{\rho}$ as detailed in [15]. As a summary of this Gaussian mixture, we can determine its mean and covariance

$$
\begin{aligned}
\boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, I) &\triangleq \sum_i \rho_i\, \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I) \\
\mathbf{C}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, I) &\triangleq \sum_i \rho_i \,\big(\mathbf{C}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I) \\
&\qquad + \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)\; \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, \theta_i,\, I)^{\mathsf{T}}\big) \\
&\qquad - \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, I)\; \boldsymbol{m}(\mathring{\boldsymbol{y}}_\star \,|\, \boldsymbol{z}_d,\, I)^{\mathsf{T}}\,.
\end{aligned}
$$

To determine the posterior distribution for hyperparameter θ_f by marginalizing over all other hyperparameters θ_{-f}, we must evaluate

$$
p(\theta_f \mid I_d) = \frac{\int p(\boldsymbol{z}_d \mid \theta, I)\, p(\theta \mid I)\,\mathrm{d}\theta_{-f}}{\int p(\boldsymbol{z}_d \mid \theta, I)\, p(\theta \mid I)\,\mathrm{d}\theta}\,. \tag{7}
$$

While these integrals are also non-analytic, Bayesian Monte Carlo again gives us a means of approximating them. Further details on the estimation of such posterior distributions can be found in [13].

The key feature of our approach is the treatment of the characteristics of changepoints and faults as covariance hyperparameters. As such, we assign to each an appropriate prior distribution, and then marginalise them using (6), 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 (7). In particular, we are able to produce full posterior distributions for the types and locations of changepoints and faults.

While the covariance functions from [13] were originally developed for single changepoints, they can be readily 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 marginalising the hyperparameters associated with the entire data stream. If no changepoint is present in the window, the posterior distribution for its location will typically be concentrated at its trailing edge. A changepoint at such a location will have no influence on predictions; the model is therefore able to effectively manage the absence of changepoints.

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 its parameters as hyperparameters of the model, and place appropriate hyperparameter samples corresponding to, for example, a constant mean value before and after a putative changepoint. The different possible mean functions will then be properly marginalised for prediction, and the likelihoods associated with the samples can give support for the proposition of a changepoint having occurred at a particular time.

III. oActive data selection in the presence of changepoints

We now turn to the central contribution of this paper, the problem of deciding which observations should be taken. Consider the decision task facing the central agent controller of the network, at an arbitrary time t. There are two components to this decision. First, we must decide whether to take an observation at all at that time. If we do, we must then select a type of observation to be made at that time—if there are multiple sensors, we must select one to draw an observation from. As previously mentioned, our goals in making this decision are likewise two-fold: we aim to both minimise the uncertainty we possess about variables of interest, and also to minimise the number of observations required to be taken. We define the objects of our interest at t as $\boldsymbol{y}_\star \triangleq \{y_l,\, l = 1,\, \ldots,\, L\}$, where y_l is the value of our field at input $[l, t + \epsilon]$, that is, at the location of sensor l for the lookahead-shifted time $t + \epsilon$. Should the uncertainty about any other points be relevant to a problem (for example, at other points we are unable to directly observe), these could also be readily incorporated into our set. We can now define the expected loss of choosing an observation from sensor l_c as

$$
\begin{aligned}
&\Lambda(l_c \mid \boldsymbol{z}_d,\, I) \\
&\triangleq \int \left(\sum_{l=1}^{L} \sqrt{\mathbf{C}(\mathring{y}_l \,|\, z_c,\, \boldsymbol{z}_d,\, I)}\right) p(z_c \mid \boldsymbol{z}_d,\, I)\,\mathrm{d}z_c + C_c\,,
\end{aligned}
\tag{8}
$$

where we have marginalised over the possible observations z_c received from sensor l_c. This integral is non-analytic, and as such will be approximated by further application of Bayesian Monte Carlo techniques.

Essentially, we treat the uncertainty associated with each variable equally, combining them all with a sum. As another simple extension, the various standard deviations could be weighted against one another if the monitoring of some variables was more critical than others.

The term C_c denotes the cost of the observation, representing a "conversion factor" between the loss associated with sensing and with uncertainty. The cost represents the average increase in the standard deviation at our pertinent locations we would require before an additional

observation would become worthwhile. It is trivially possible (both theoretically and computationally) to take a different C_c for different types of observation. However, our applications in this paper require only a fixed cost C, and therefore we consider only this possibility.

We also have the loss associated with not taking an observation at all

$$\Lambda(\emptyset \mid \boldsymbol{z}_d,\, I) \triangleq \sum_{l=1}^{L} \sqrt{\mathbf{C}(\mathring{y}_l \,|\, \boldsymbol{z}_d,\, I)}\,. \tag{9}$$

So, defining

$$l_m \triangleq \underset{l_c=1,\,\ldots,\,L}{\operatorname{argmin}} \; \Lambda\left(l_c \mid \boldsymbol{z}_d,\, I\right)\,,$$

if $\Lambda\left(l_m \mid \boldsymbol{z}_d,\, I\right) < \Lambda(\emptyset \mid \boldsymbol{z}_d,\, I)$, we sample at l_m; otherwise, we do not sample at all. Of course, it is usually not desirable to have to evaluate this policy at every time it is possible to sample. Instead, after taking an observation, we can use any optimisation algorithm to determine the future time at which a sample will next become necessary.

Importantly, the variances in (8) and (9) capture our underlying uncertainty about the correct model, due to our principled marginalisation. For example, the uncertainty associated with a sensor being faulty, and the potential characteristics of that fault, all influence the variances we are trying to minimise. More appropriate observations will be scheduled if there is a significant degree of uncertainty about the model. Note also that while we make a decision at time t solely with the objective of minimising our uncertainty about the state at $t + \epsilon$, the smoothness of a GP means that our scheduled observations will also give low uncertainty for subsequent times.

It is worth noting the differences between the loss function described above, and that used in previous approaches. In [15], a loss function was taken that was infinite if the variance associated with any object of interest become greater than a pre-specified threshold. This is a problematic choice when our sensors are subject to faults. Should a sensor become faulty, the uncertainty associated with it is likely to increase beyond the threshold, regardless of the samples we take. Once this has happened, the algorithm will request observations from it constantly—clearly, this is undesirable behaviour.

IV. Results

We have applied our methods to several datasets. Specifically, we perform prediction for a variable using (6), and use those predictions in order to perform active data selection as per Section III. Although we do not have active control over the sensors in question, all datasets were sufficiently dense that we can select from the pre-recorded observations without being significantly constrained by the available times of observation. The full list of all available observations additionally serves as a measure of "ground truth" in order to verify the accuracy of our predictions, except, of course, where the observations are faulty. Finally, we produce a posterior over changepoint/fault locations by estimating (7), as required.

A. Bramblemet weather sensor network

The first test of our algorithm was performed on a network of weather sensors located on the south coast of England[2]. In particular, 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 www.sotonmet.co.uk). This sensor is subject to network outages and other faults that suggest the use of our approach.

Figure 2 demonstrates faulty data selection over a dataset featuring a faulty period in which the observations returned by the sensor become stuck at a reading of 4m. Float sensors, such as are often used to measure tides, are prone to becoming lodged on sensor posts, giving rise to this kind of behaviour. As such, we used a model that allowed for changes in the observation likelihood of this type. The covariance function used for non-faulty behaviour was that used for tide heights in [15], with its hyperparameters (such as tidal period and amplitude) determined off-line from the large corpus of such data. Hence we were required to marginalise only the hyperparameters that directly specify the fault[3]. This necessitated the stipulation of a prior over the usual length of such faults. We took a prior over the natural logarithm of this length with a mean corresponding to a fault length of 12 hours and a standard deviation of 0.5. A grid of hyperparameter samples was taken with 11 samples in the length of the fault, and 100 samples in its location. The cost of an observation was taken to be 0.075m.

Our algorithm correctly detects the stuck fault beginning at around $t = 0.8$ days, reducing its sampling frequency as a result. More observations are selected as it begins to suspect that the fault may have ended. Accurate predictions, with realistic uncertainties, are made throughout.

B. Wannengrat weather sensor network

Our second dataset was drawn from the Wannengrat Alpine Observatory being built above and around the town of Davos as part of Swiss Experiment (see www.swiss-experiment.ch/index.php/Wannengrat:Home). This comprises a wireless sensor network deployed with the aim of understanding the complex metereological processes occurring on the snowpack. The changepoints that exist within this dataset, along with the limited battery life of the remote sensors, make it a natural fit for our methods. In particular, we perform active data sampling over a dataset comprising ambient temperature measurements featuring a dramatic changepoint.

We express the covariance over sensors $K^{(l)}$ as a function of the (isotropic) spatial distance between the known

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

[3]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.

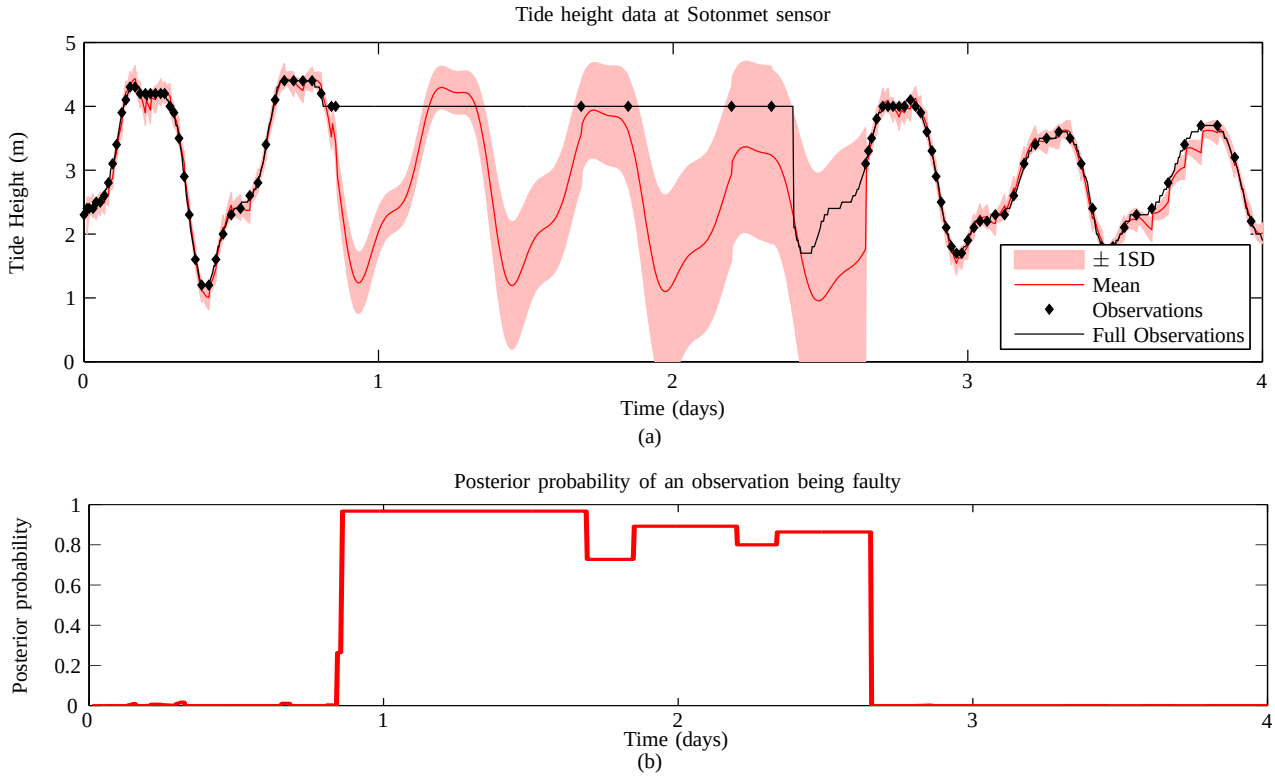

Figure 2: Active data selection over intermittently faulty tide height data from the Sotonmet sensor—(a) the selected observations and consequent predictions (with lookahead $\epsilon = 5$ mins), (b) the posterior probability that an observation taken at a time $t + \epsilon$ would be faulty, given all observations taken up to and including time t.

sensor locations. We used the Matérn covariance function (2) for this purpose. The covariance over time $K^{(t)}$ was another Matérn covariance of the form (2), modified to allow for changepoints in output scale, using $K^{(\mathrm{E})}$ from Figure 1c. Our grid of hypersamples had 50 samples in the location of the changepoint, 7 samples in both the output scale before and after a putative changepoint, 5 samples in the input scale and 3 samples in the noise variance η. The cost of observation was taken to be $0.5°$C. Although slightly higher than is realistic given the actual batteries used, this cost did allow us to produce visually intelligible plots.

Figure 3 demonstrates active data selection over this dataset. Two initial samples are taken from each sensor at near-identical times

C. EEG data with saccade event

EEG is a non-invasive measure of brain activity. However, the EEG signal is often corrupted by eye movement artifacts referred to as saccades; it is necessary to remove the artifact from the EEG signal. We treat such saccade events as faults, and use our method to select observations in this context.

Effectively, the observations are the superposition of the EEG signal and a fault signal, requiring a simple modification of our observation likelihood [17]. A typical fault signal during saccade (depicted in Fig. 4) is a function that undergoes a temporary excursion from an otherwise constant value of zero. As such, it is modelled using an appropriate 'drift' covariance. We used a prior upon the logarithm of the saccade duration that was a Gaussian with a mean of log(110ms) and a standard deviation of 0.6. Using a grid of hyperparameter samples, we took 100 samples in the location of the fault, 9 in its length and 5 for the scale of variation of the drift. The cost of observation used was 0.1.

Fig. 5 displays active data selection and prediction over our EEG data, demonstrating the identification of the fault. Sampling is reduced during the fault, when observations become more (but not completely) uninformative.

V. Conclusion

We have introduced a new sequential algorithm for performing active data selection in the presence of changepoints or faults. We employ a principled, Bayesian framework throughout, performing prediction using a Gaussian process with an appropriate changepoint covariance. We then used those predictions to evaluate the optimal decision-theoretic sampling policy, given the goal of minimising the average uncertainty for a defined cost of taking an observation. Tests on real sensor networks demonstrate the efficacy of our approach.

Acknowledgment

This research was partly funded by the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Networks) project and the SEAS DTC project AA021.

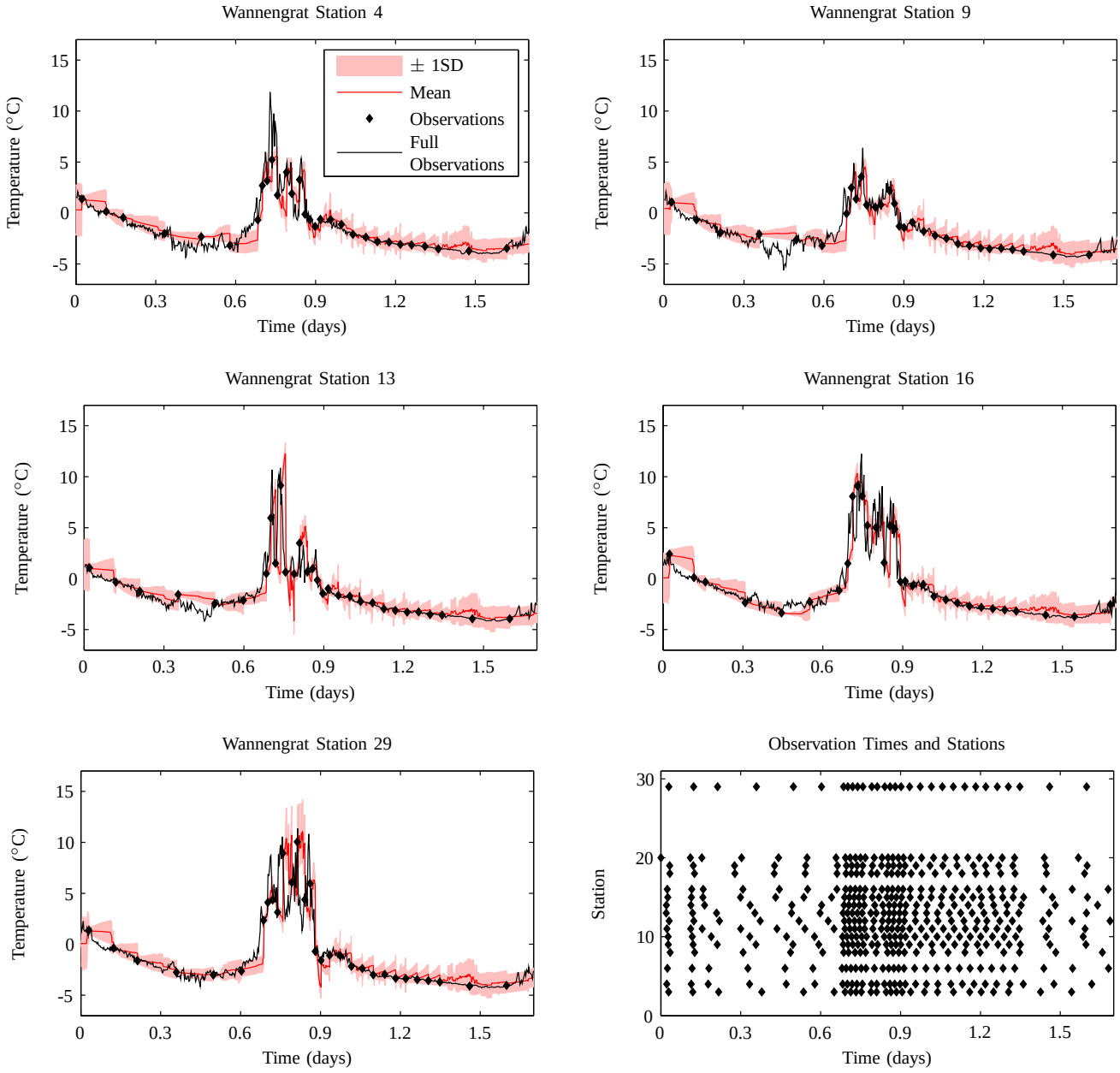

Figure 3: Active data selection of ambient temperatures at 16 Wannengrat sensor stations. Displayed predictions were made with zero lookahead, $\epsilon = 0$.

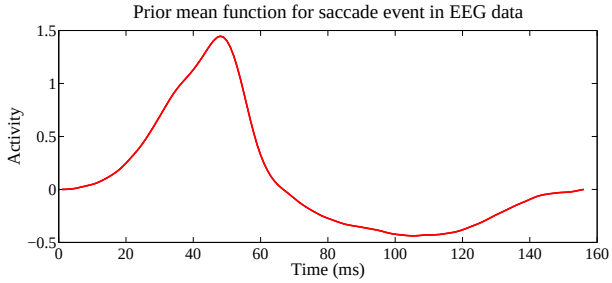

Figure 4: Eye movement activity during a saccade event.

REFERENCES

[1] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.

[2] S. Seo, M. Wallat, T. Graepel, and K. Obermayer, "Gaussian process regression: active data selection and test pointrejection," in *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 3, 2000.

[3] M. Basseville and I. Nikiforov, *Detection of abrupt changes: theory and application*. Prentice Hall, 1993.

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

[5] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, p. 711, 1995.

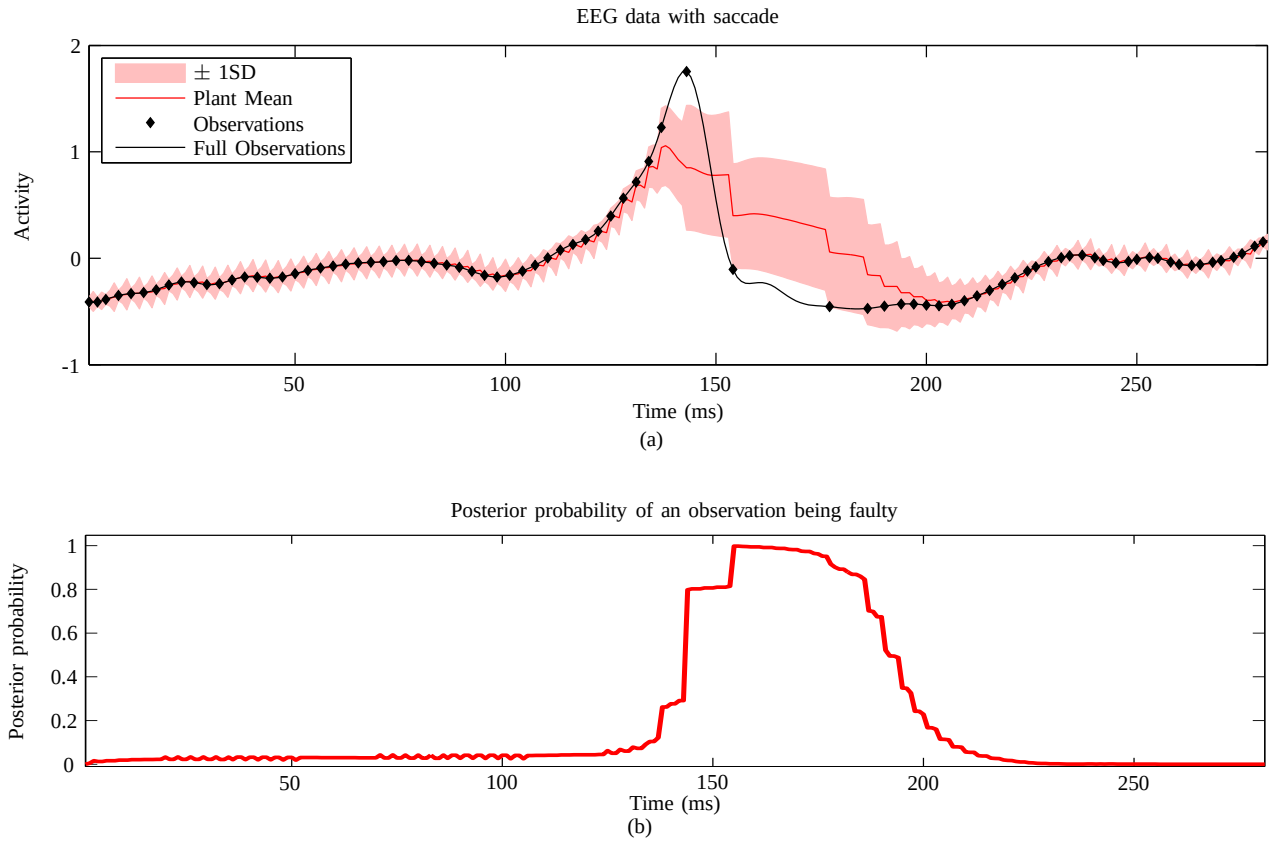

Figure 5: Active data selection over EEG data containing a saccade fault – (a) the selected observations and consequent tracking predictions (with zero lookahead), (b) the posterior probability that an observation taken at a time t would be faulty, given all observations taken up to and including time t.

[6] M. Csorgo and L. Horvath, *Limit theorems in change-point analysis*. John Wiley & Sons, 1997.

[7] J. Chen and A. Gupta, *Parametric Statistical Change Point Analysis*. Birkháuser Verlag, 2000.

[8] R. P. Adams and D. J. MacKay, "Bayesian online changepoint detection," University of Cambridge, Cambridge, UK, Tech. Rep., 2007, arXiv:0710.3742v1 [stat.ML].

[9] P. Fearnhead and Z. Liu, "On-line inference for multiple changepoint problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 589–605, 2007.

[10] T. Kobayashi and D. Simon, "Application of a bank of Kalman filters for aircraft engine fault diagnosis," in *Proceedings of ASME Turbo Expo 2003, Power for Land, Sea, and Air, June 16-19 2003, Atlanta, Georgia, USA*, 2003.

[11] V. Aggarwal, K. Nagarajan, and K. Slatton, "Estimating failure modes using a multiple-model Kalman filter," ASPL, Tech. Rep. no. Rep_2004-03-001, 2004. [Online]. Available: http://www.aspl.ece.ufl.edu/reports/Rep_2004-03-001.pdf

[12] S. Reece, C. Claxton, D. Nicholson, and S. J. Roberts, "Multi-Sensor Fault Recovery in the Presence of Known and Unknown Fault Types," in *Proceedings of the 12th International Conference on Information Fusion (FUSION 2009), Seattle, USA*, 2009.

[13] R. Garnett, M. A. Osborne, and S. Roberts, "Sequential Bayesian prediction in the presence of changepoints," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM New York, NY, USA, 2009.

[14] S. Reece, R. Garnett, M. A. Osborne, and S. J. Roberts, "Anomaly detection and removal using non-stationary Gaussian processes," University of Oxford, Oxford, UK, Tech. Rep., 2009. [Online]. Available: http://www.robots.ox.ac.uk/~reece/reece2009.pdf

[15] M. A. Osborne, A. Rogers, S. Ramchurn, S. J. Roberts, and N. R. Jennings, "Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes," in *International Conference on Information Processing in Sensor Networks 2008*, April 2008, pp. 109–120. [Online]. Available: http://eprints.ecs.soton.ac.uk/15122/

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

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

[18] C. E. Rasmussen and Z. Ghahramani, "Bayesian Monte Carlo," in *Advances in Neural Information Processing Systems*, S. Becker and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, vol. 15.