Innovative And Additive Outlier Robust Kalman Filtering With A Robust Particle Filter

Alexander T. M. Fisch, Idris A. Eckley, Paul Fearnhead

***Abstract*—In this paper, we propose CE-BASS, a particle mixture Kalman filter which is robust to both innovative and additive outliers, and able to fully capture multi-modality in the distribution of the hidden state. Furthermore, the particle sampling approach re-samples past states, which enables CE-BASS to handle innovative outliers which are not immediately visible in the observations, such as trend changes. The filter is computationally efficient as we derive new, accurate approximations to the optimal proposal distributions for the particles. The proposed algorithm is shown to compare well with existing approaches and is applied to both machine temperature and server data.**

***Index Terms*—Kalman Filter, Anomaly Detection, Particle Filtering, Robust Filtering**

I. Introduction And Literature Review

Anomaly detection is an area of considerable importance and has been subject to increasing attention in recent years. Comprehensive reviews of the area can be found in [1, 2]. The field's growing importance arises from the increasing range of applications to which anomaly detection lends itself: from fraud prevention [1, 2], to fault detection [1, 2], and even the detection of exoplanets [3]. More recently, the emergence of the internet of things and the ubiquity of sensors has led to emergence of the online detection of anomalies as an important statistical challenge.

Kalman filters [4] provide a convenient framework to detect anomalies within a streaming data context. In particular, they can be updated in a fully online fashion at a fixed computational cost. At each time point, Kalman filters also provide an estimate both for the expectation and variance of the next observation. These can be used to determine whether that observation is anomalous or not. However, the major drawback of Kalman filters is their lack of robustness to outliers: once the filter has encountered an outlier, it will often produce inaccurate predictions for many future time points.

The anomaly detection literature distinguishes between two types of outliers. The first are additive outliers, sometimes referred to as observational outliers [5], which affect the observational noise only. The other type are innovative, or process [6], outliers. These affect the updates of the hidden states. In practice, both have a similar effect on the next observation, but quite different effects on subsequent observations. Moreover, some innovative outliers cannot be detected immediately as their influence on the observations is only noticeable after, or over, a period of time.

A range of robust Kalman filters has been proposed to date. Many side-step the problem of distinguishing between the two outlier types. By far the largest class of filters aims to be robust against heavy tailed additive outliers. Examples of such filters include [7, 8], which assume t-distributed additive noise and perform inference using variational Bayes, [9], who use Huberised, i.e. truncated, residuals, and [10] who inflate the noise covariance matrix whenever an outlier is encountered. A few filters have also been developed with the aim of achieving robustness against innovative outliers [9]. The problem with such filters is that they exacerbate the shortcomings of the Kalman filter when they encounter the other type of anomaly: additive outlier robust Kalman filters, for example, update their hidden states even less than the classical Kalman filter when encountering innovative outliers.

In principle, it seems straightforward to combine the ideas of these two types of robust Kalman filter. One body of literature proposes to use Huberisation of both innovative and additive residuals [5, 10]. Others [6, 11] have modelled both additive and innovative outliers using t-distributions, by imposing Wishart priors on the precision matrix of both the innovations and additions and maintaining the posterior by using variational Bayes approaches. The issue with these filters comes from how they approximate the filtering distribution of the state. Both return uni-modal posteriors after encountering an anomaly. This is a shortcoming given that the posterior after an anomaly is likely to be multi-modal (see Figure 2 below) as different types of anomalies contain different amounts of information about the state: If we have an anomaly at time t, then if this is an additive anomaly it has little information about the state at time t, and thus the new filtering distribution for the state will be close to the predictive distribution for the state given the data up to time $t-1$. Whereas if it is an innovative anomaly then the state will have changed substantially from what was predicted.

The ideal approach to constructing a robust filter would be to model the possibility of outliers in both the observation and system noise, and then use a filter algorithm that attempts to calculate, or approximate, the true filtering distribution for the model. An early attempt to do this was the spline based approach of [12], but the computational complexity increases very quickly with the number of dimensions and such a filter becomes impracticable when the state dimension is greater than 3. As a result we consider using particle filters [13, 14]. These are able to produce Monte Carlo approximations to the filtering distribution for an appropriate model that allows for outliers, and, in principle, can work even if the filtering distribution is multi-modal. However the Monte Carlo error of standard implementations of the particle filter can be prohibitively large [10].

In this paper, we develop an efficient particle filter by using

a combination of Rao-Blackwellisation and well-designed proposal distributions. The idea of Rao-Blackwellisation is to integrate out part of the state so that the particle filter approximates the filtering distribution of a lower-dimensional projection of the state. In our application this projection is whether each component of the additive and innovative noise is an outlier, and if it is how much the variance of the noise has been inflated. Conditional on this information, the state space model becomes linear-Gaussian and we can implement a Kalman Filter to calculate exactly the conditional filtering distribution, while being able to fully capture multi modal posteriors. This idea is similar to that which underpins the Mixture Kalman Filter [15].

Whilst Rao-Blackwellisation improves the Monte Carlo accuracy of the filter, such a filter can still have the shortcomings noted by [10] and perform poorly without good proposal distributions for the information we condition on. One of the main contributions of this work is a proposal distribution that accurately approximates the conditional distribution of the variance inflation for each component of the noise, and hence approximates the optimal proposal distribution [16]. As a result of this proposal, we find that accurate results can be obtained even with only a few particles.

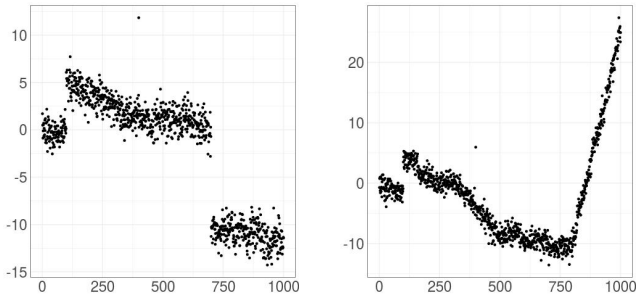
Another important challenge addressed by this paper is that certain innovative outliers can not immediately be detected. An innovative outlier in a latent trend component for instance can cause a trend change which may only become apparent – i.e. produce a visible outlier in the observations – many observations after the innovative outlier in the trend occurred. It is nevertheless important to capture such outliers as they can affect a potentially unlimited number of observations to come. The proposed particle filter includes the possibility to back-sample the variance inflation particles in light of more recent observations, which enables it to capture these important anomalies.

The remainder of this paper is organised as follows: We discuss our robust noise model, consisting of a mixture distribution of Gaussian noise, representing typical behaviour, and heavy tailed noise, representing atypical behaviour, for both the additive (observational) and innovative (system) noise process in Section II. We then introduce the proposal distribution for the scale of the noise in Section III, before extending it to anomalies which are not immediately identifiable in Section IV. The proposed filter is compared to others in Section V and applied to router data and a benchmark machine temperature data-set in Section VI. The proposed methodology, which we call Computationally Efficient Bayesian Anomaly detection by Sequential Sampling (CE-BASS) has been implemented in the the `R` package `RobKF` available on CRAN [17].

II. Model And Examples

Throughout this paper, we will consider inference about a latent state, $\mathbf{X}_t$, through partial observations, $\mathbf{Y}_t$, modelled as

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{C}\mathbf{X}_t + \mathbf{\Sigma}_A^{\frac{1}{2}}\mathbf{V}_t^{\frac{1}{2}}\boldsymbol{\epsilon}_t, \\ \mathbf{X}_t &= \mathbf{A}\mathbf{X}_{t-1} + \mathbf{\Sigma}_I^{\frac{1}{2}}\mathbf{W}_t^{\frac{1}{2}}\boldsymbol{\nu}_t. \end{aligned} \tag{1}$$



(a) Random walk (b) Random walk with trends

Fig. 1: Two examples of time series which are realisations of outlier infested Kalman models. (a) was simulated using the setup defined in Equation (2), with $\sigma_A = 1$, $\sigma_I = 0.1$, and outliers defined by $W_{100} = 3600$, $V_{400} = 100$, and $W_{700} = 10000$. Conversely (b) second example was simulated using the model defined in Equation (3) using $\sigma_A = 1$, $\sigma_I^{(1)} = 0.1$, $\sigma_I^{(2)} = 0.01$ and outliers defined by $W_{100}^{(1)} = 3600$, $V_{400} = 100$, and $W_{700}^{(2)} = 40000$.

Here the additive noise, $\boldsymbol{\epsilon}_t \in \mathbb{R}^p$, and the innovations $\boldsymbol{\nu}_t \in \mathbb{R}^q$ are both i.i.d. standard multivariate Gaussian. The matrices $\mathbf{\Sigma}_A$ and $\mathbf{\Sigma}_I$ denote the covariance of the additive and innovation noise respectively. Without loss of generality we assume that these matrices are diagonal, as a general model can be transformed to one which satisfied this assumption by applying a suitable rotation to the observation and/or the state (see Section III in the Supplementary Material for details). The diagonal matrices $\mathbf{V}_t$ and $\mathbf{W}_t$ are used to capture additive and innovative outliers respectively, with large diagonal entries of $\mathbf{V}_t$ corresponding to additive outliers and large diagonal entries of $\mathbf{W}_t$ corresponding to innovative outliers. The classical Kalman model is recovered by setting $\mathbf{W}_t = \mathbf{I}$ and $\mathbf{V}_t = \mathbf{I}$ for all times t.

The model in Equation (1) can be used to model a range of time series behaviours. We will use the following two examples throughout the paper:

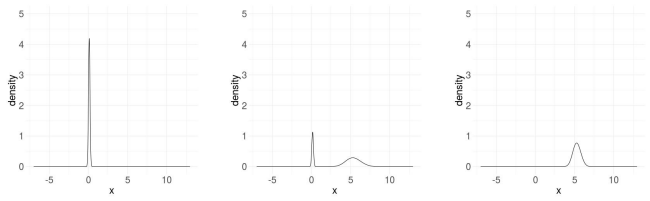
Example 1: The random walk model with both change-points and outliers, similar to the problem considered by [18]. It can be formulated as

$$Y_t = X_t + V_t^{\frac{1}{2}}\sigma_A\epsilon_t, \qquad X_t = X_{t-1} + W_t^{\frac{1}{2}}\sigma_I\nu_t. \tag{2}$$

Here atypically large values of V_t correspond to outliers, whilst atypically large values of W_t correspond to changes. A realisation of this model can be found in Figure 1a. This example illustrates the challenge of the bi-modal hidden state distribution introduced by anomalies. Figure 2 expands on this point.

Example 2: A time series with changes in trend, level shifts, as well as outliers, similar to the model considered by [19]. It can be formulated as

$$\begin{aligned} Y_t &= X_t^{(1)} + V_t^{\frac{1}{2}}\sigma_A\epsilon_t \\ X_t^{(1)} &= X_{t-1}^{(1)} + X_{t-1}^{(2)} + \left(W_t^{(1)}\right)^{\frac{1}{2}}\sigma_I^{(1)}\nu_t^{(1)}, \\ X_t^{(2)} &= X_{t-1}^{(2)} + \left(W_t^{(2)}\right)^{\frac{1}{2}}\sigma_I^{(2)}\nu_t^{(2)}, \end{aligned} \tag{3}$$



(a) $x_{99}|Y_{0:99}$ (b) $x_{100}|Y_{0:100}$ (c) $x_{101}|Y_{0:101}$

Fig. 2: The distribution of the hidden state x_t for the process depicted in Figure 1a. When we observe the abrupt change in the observations at time 100, we have a bi-modal posterior as the observation may be an additive or an innovative outlier.

with the first component of the hidden state denoting the current position and the second indicating the trend. Here, outliers are modelled by large values of V_t whilst level shift and changes in trend are modelled by atypically large values of $W_t^{(1)}$ and $W_t^{(2)}$ respectively. A realisation of this model can be found in Figure 1b.

A key feature of this second model is that an outlier in the trend component, $X_t^{(2)}$, may only become detectable many observations after the outlier – this challenging issue mentioned in the introduction is addressed via the methods in Section IV. A wide rage of other commonly used time series features, such as auto-correlation, moving averages, etc. can be incorporated in the model.

In the rest of the paper we will use the notation that a superscript (i, j) on a matrix refers to the (i, j)th entry of that matrix, and the superscript $(i, :)$ refers to the ith row. To infer the locations of anomalies we use the model

$$\mathbf{V}_t^{(i,i)} = 1 + \lambda_t^{(i)} \frac{1}{\tilde{\mathbf{V}}_t^{(i,i)}} \qquad \mathbf{W}_t^{(j,j)} = 1 + \gamma_t^{(j)} \frac{1}{\tilde{\mathbf{W}}_t^{(j,j)}} \quad (4)$$

for $1 \leq i \leq p$ and $1 \leq j \leq q$. The Bernoulli random variables $\lambda_t^{(i)} \sim Ber(r_i)$ and $\gamma_t^{(j)} \sim Ber(s_j)$ are indicators that determine whether an anomaly is present or not for $1 \leq i \leq p$ and $1 \leq j \leq q$ respectively. For additional interpretability, we impose that at most one anomaly is present at any given time t, and define r_i and s_j to be the probabilities that $\lambda_t^{(i)} = 1$ and $\gamma_t^{(j)} = 1$ respectively. The inverse scale, or precision, of an anomaly is assumed to be distributed as a scaled gamma random variable. That is if $\Gamma(a, b)$ denotes a gamma random variable with shape parameter a and rate parameter b, then $\tilde{\mathbf{V}}_t^{(i,i)} \sim \tilde{\sigma}_i \Gamma(a_i, a_i)$ and $\tilde{\mathbf{W}}_t^{(j,j)} \sim \hat{\sigma}_j \Gamma(b_j, b_j)$ for $1 \leq i \leq p$ and $1 \leq j \leq q$ respectively.

The proposed model bears similarities to the model used by [11]. Both use a mixture of Gaussian and heavy tailed noise. The main difference is that the anomalous behaviour is characterised by noise which is the sum of a Gaussian and a t-distribution in our model as opposed to just a t-distribution in the model used by [11]. This ensures that anomalies coincide with strictly greater noise and makes the result more interpretable. In practice, however, the noise distribution considered in this paper and in [11] are likely to be of very similar shape.

III. Particle Filter

We now turn to filtering the model defined by Equations (1) and (4). The main feature we exploit is the fact that if we knew the value of $(\mathbf{V}_t, \mathbf{W}_t)$ at all times t, we could just run the classical Kalman filter over the data. Consequently, our approach will consist of sampling particles for $(\mathbf{V}_t, \mathbf{W}_t)$, conditional on which the classical Kalman update equations for the hidden state $\mathbf{x}_t$ can be used. This approach, very similar to the mixture Kalman filter [15, 20], is summarised by the pseudocode in Algorithm 1. Details of sub-routines for this and later algorithms can be found in Section VI of the Supplementary Material.

For each time, t, the code loops over the existing particles, $(\mathbf{V}_t, \mathbf{W}_t)$, and simulates M' descendants for each of them in Step 4. They and their associated weights, denoted by $prob$, are stored in a set of candidate particles. If we have N particles at time t, keeping all candidates would produce NM' particles at time $t+1$. To avoid the number of particles growing exponentially with t, Step 7 resamples the candidates with probability proportional to their weights to keep just N particles; there are various algorithms that can be used, see [21]. The filtering distribution for each of these particles is then calculated using the Kalman Filter updates in Step 10. As the particles store the $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ matrices it is simple to extract information about whether there have been any outliers: if a particle has an entry on the diagonal of $\boldsymbol{V}_t$ or $\boldsymbol{W}_t$, that is not one then that particle corresponds, respectively, to an additive or innovative outlier.

Algorithm 1 Basic Particle Filter (No Back-sampling)

Input: An initial state estimate $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
A number of descendants, $M' = M(p+q)+1$
A number of particles to be maintained, N.
A stream of observations $\mathbf{Y}_1, \mathbf{Y}_2, ...$
Initialise: Set *Particles*$(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\}$
1: **for** $t \in \mathbb{N}$ **do**
2: *Candidates* $\leftarrow \{\}$
3: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in$ *Particles*(t) **do**
4: $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_Particles$(M', \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_{t+1}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
5: *Candidates* $\leftarrow$ *Candidates* $\cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob)\}$
6: **end for**
7: *Descendants* $\leftarrow$ Resample$(N,$ *Candidates*$)$
8: *Particles*$(t+1) \leftarrow \{\}$
9: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob) \in$ *Descendants* **do**
10: $(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+1}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}\boldsymbol{\Sigma}_A, \mathbf{W}\boldsymbol{\Sigma}_I)$
11: *Particles*$(t+1) \leftarrow$ *Particles*$(t+1) \cup \{(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})\}$
12: **end for**
13: **end for**

The main challenge in the above approach consists of selecting a good sampling procedure for the particles. Whilst it may be a natural choice to sample particles $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$ from their prior distribution, this is not suitable for the problem considered in this paper. In particular, this sampling procedure would not be robust to outliers: the stronger an anomaly was, the less likely we would be to sample a particle with an appropriate value of $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$, as discussed by [10].

Adopting ideas from [16] and [22], we overcome the above challenge by sampling particles from an approximation to the conditional distribution of $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$ given observation $\mathbf{Y}_{t+1}$. Denote the model's prior distribution for $(\mathbf{V}_{t+1}, \mathbf{W}_{t+1})$ in (4) by $\pi_0(\cdot)$. The conditional distribution

$\pi(\mathbf{W}_{t+1}, \mathbf{V}_{t+1}|\mathbf{Y}_{t+1})$ for the descendants of a particle whose filtering distribution for $\mathbf{x}_t$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is then proportional to

$$\pi_0(\mathbf{W}, \mathbf{V})\mathcal{L}\left(\mathbf{Y}, \mathbf{CA}\boldsymbol{\mu}, \mathbf{CA}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{C}^T + \boldsymbol{\Sigma}_A\mathbf{V} + \mathbf{C}\boldsymbol{\Sigma}_I\mathbf{W}\mathbf{C}^T\right).$$

Here we have dropped time indices for convenience, and $\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the likelihood of an observation $\mathbf{x}$ under a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-model. Since at most one component is anomalous, we can re-write this as a sum over which, if any, component is anomalous

$$\mathbb{I}_{\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}\}}\pi(\mathbf{I}, \mathbf{I}|\mathbf{Y}) + \sum_{j=1}^{q}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I}+\frac{\mathbf{I}^{(j)}}{\tilde{\mathbf{W}}^{(j,j)}},\mathbf{V}=\mathbf{I}\right\}}\hat{\pi}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right) + \sum_{i=1}^{p}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}+\frac{\mathbf{I}^{(i)}}{\tilde{\mathbf{V}}^{(i,i)}}\right\}}\tilde{\pi}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right).$$

Here, $\mathbf{I}^{(j)}$ denote a matrix whose (j,j)th entry is 1 and all other entries are zero, and we use the shorthand

$$\tilde{\pi}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right) = \pi\left(\mathbf{I}, \mathbf{I} + \frac{\mathbf{I}^{(i)}}{\tilde{\mathbf{V}}^{(i,i)}}|\mathbf{Y}\right)$$

and

$$\hat{\pi}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right) = \pi\left(\mathbf{I} + \frac{\mathbf{I}^{(j)}}{\tilde{\mathbf{W}}^{(j,j)}}, \mathbf{I}|\mathbf{Y}\right).$$

Since the target distribution $\pi(\mathbf{W}, \mathbf{V}|\mathbf{Y})$ is intractable, we construct an approximation to it, which we denote $q(\mathbf{W}, \mathbf{V}|\mathbf{Y})$, and use this as our proposal distribution. This proposal is proportional to

$$\mathbb{I}_{\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}\}}\beta_0 + \sum_{j=1}^{q}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I}+\frac{\mathbf{I}^{(j)}}{\tilde{\mathbf{W}}^{(j,j)}},\mathbf{V}=\mathbf{I}\right\}}\hat{\beta}_j\hat{q}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right) + \sum_{i=1}^{p}\mathbb{I}_{\left\{\mathbf{W}=\mathbf{I},\mathbf{V}=\mathbf{I}+\frac{\mathbf{I}^{(i)}}{\tilde{\mathbf{V}}^{(i,i)}}\right\}}\tilde{\beta}_i\tilde{q}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right).$$

Clearly, there is no benefit in simulating multiple identical descendants, so we wish to sample precisely one dependent that corresponds to no outliers. To do this, and also to have the same number of descendant particles for each possible type of outlier, we set $\beta_0 = \frac{1}{1+M(p+q)}$, $\tilde{\beta}_i = \frac{M}{1+M(p+q)}$, and $\hat{\beta}_j = \frac{M}{1+M(p+q)}$, and use stratified subsampling as in [20]. This leads to $M' = M(p+q)+1$ total descendants per particle, M for each of the p additive and q innovative outliers, and one for no outlier. Each of these particles is then given a weight proportional to

$$\frac{\pi(\mathbf{W}_{t+1}, \mathbf{V}_{t+1}|\mathbf{Y}_{t+1})}{q(\mathbf{W}_{t+1}, \mathbf{V}_{t+1}|\mathbf{Y}_{t+1})}.$$

The main challenge now consists of obtaining proposal distributions $\tilde{q}_i(\cdot)$ for $1 \leq i \leq p$ and $\hat{q}_j(\cdot)$ for $1 \leq j \leq q$ that provide good approximations to the conditional posteriors which are proportional to $\tilde{\pi}_i(\cdot)$ and $\hat{\pi}_j(\cdot)$ respectively. In the next subsection, we therefore derive proposal distributions that provide leading order approximations to the conditional posteriors. To simplify notation, we define the predictive variance $\hat{\boldsymbol{\Sigma}} = \mathbf{CA}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{C}^T + \boldsymbol{\Sigma}_A + \mathbf{C}\boldsymbol{\Sigma}_I\mathbf{C}^T$ and use it throughout the remainder of this paper. We also begin by assuming that $\mathbf{C}$ contains no columns that are identically 0, as if this is the case then the observation at time t contains no information about at least one component of the state at time t. The proposal introduced in the following subsection also forms the basis of back-sampling introduced in Section IV, which allows us to relax this assumption on $\mathbf{C}$.

A. Proposal Distributions

For $1 \leq i \leq p$, we would like the proposal distribution $\tilde{q}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right)$ for the precision, $\tilde{\mathbf{V}}^{(i,i)}$, to be as close as possible to $\tilde{\pi}_i\left(\tilde{\mathbf{V}}^{(i,i)}\right)$ or, equivalently, proportional to

$$f_i\left(\tilde{\mathbf{V}}^{(i,i)}\right)\frac{\exp\left(-\frac{1}{2}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})^T\left(\hat{\boldsymbol{\Sigma}} + \frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})\right)}{\sqrt{\left|\hat{\boldsymbol{\Sigma}} + \frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right|}},$$

where $f_i()$ denotes the PDF of the $\tilde{\sigma}_i\Gamma(a_i, a_i)$-distributed prior of $\tilde{\mathbf{V}}^{(i,i)}$.

It should be noted that the intractable terms,

$$\left|\hat{\boldsymbol{\Sigma}} + \frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right| \quad \text{and} \quad \left(\hat{\boldsymbol{\Sigma}} + \frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1} \tag{5}$$

can both be expanded using the matrix determinant lemma and the Sherman-Morrison formula respectively, as they are rank 1 updates of a determinant and inverse respectively. Indeed, by the matrix determinant lemma,

$$\left|\hat{\boldsymbol{\Sigma}} + \frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}^{(i,i)}}\mathbf{I}^{(i)}\right| = \frac{\left|\hat{\boldsymbol{\Sigma}}\right|}{\tilde{\mathbf{V}}^{(i,i)}}\left(1 + \boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} + O\left(\tilde{\mathbf{V}}^{(i,i)}\right)\right),$$

the leading order term is conjugate to the prior of $\tilde{\mathbf{V}}^{(i,i)}$. Moreover, by the Sherman Morrison formula the second term in Equation (5) is equal to

$$\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)}\hat{\boldsymbol{\Sigma}}^{-1}\left[\frac{1}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}} - \left(\frac{1}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\frac{\tilde{\mathbf{V}}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}}\right],$$

up to $O\left((\tilde{\mathbf{V}}^{(i,i)})^2\right)$. Crucially, the first two terms are constant in $\tilde{\mathbf{V}}^{(i,i)}$, while the third is linear in $\tilde{\mathbf{V}}^{(i,i)}$ and therefore returns a term which is conjugate to the prior of $\tilde{\mathbf{V}}^{(i,i)}$. Furthermore, we are most concerned about accurately sampling the particle when an anomaly occurs in the ith component, which happens when the precision, $\tilde{\mathbf{V}}^{(i,i)}$, and the higher order terms, become small.

Keeping only the leading order terms in the determinant and the exponential term results in a proposal distribution for $\tilde{\mathbf{V}}^{(i,i)}$ of the form

$$\tilde{\mathbf{V}}^{(i,i)} \sim \tilde{\sigma}_i\Gamma\left(a_i + \frac{1}{2}, a_i + \frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right).$$

More detailed derivations, including the associated weight, are given by Theorem 1 in the Supplementary Material. This proposal has the property that as the observed anomaly in the ith component becomes larger, i.e. as

$$\frac{1}{\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y}-\mathbf{CA}\boldsymbol{\mu})}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2$$

increases, the mean of the proposal for $\tilde{\mathbf{V}}^{(i,i)}$ diverges from the prior mean and behaves asymptotically like

$$(2a_i+1)\boldsymbol{\Sigma}_A^{(i,i)}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu})}\right)^2.$$

Consequently, the variance and the squared residual will be on the same scale, thus achieving computational robustness.

A very similar approach can be used to obtain a proposal distribution $\hat{q}_j\left(\tilde{\mathbf{W}}^{(j,j)}\right)$ which provides a leading order approximation for the distribution proportional to $\pi\left(\mathbf{I}+\frac{1}{\tilde{\mathbf{W}}^{(j,j)}}\mathbf{I}^{(j)},\mathbf{I}|\mathbf{Y}\right)$. The proposal consists of sampling

$$\tilde{\mathbf{W}}^{(j,j)}\sim\hat{\sigma_j}\Gamma\left(b_j+\frac{1}{2},b_j+\frac{\hat{\sigma_i}}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu}\right)}{\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2\right)$$

and is of very similar form to the proposal distribution for particles with an additive outlier and well defined if $\mathbf{C}$ has no columns that just contain zeros. Further details, including the associated weight, are given in Theorem 2 in the Supplementary Material. Like the proposal distribution for particles with an additive anomaly this proposal is computationally robust: it ensures that the squared residual and the variance will be on the same scale as the anomaly in the jth innovative component becomes stronger.

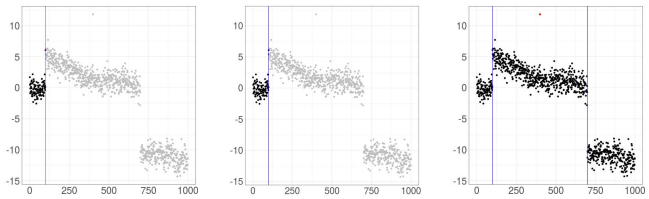
Finally, the "proposal" for particles without anomalies consists of deterministically setting $\mathbf{V}=\mathbf{I}$ and $\mathbf{W}=\mathbf{I}$. The weight associated with this particle is proportional to the likelihood, the closed form of which is given in Theorem 3 in the Supplementary Material.

B. Choices of Parameters

The choice of hyper-parameters, particularly $\hat{\sigma_i}$ and $\tilde{\sigma_i}$, has a significant effect on the performance of the proposed filter. One reason for this is that an outlier observation could be the result of either an additive or an innovative outlier. It may be that the root cause can only be determined after further observations are made. Thus, we wish to choose hyper-parameters in such a way as to ensure that observed anomalies, which are equally well explained by different classes of anomalies, are given similar importance weights. This will not automatically happen for larger outlier observations, as the model could asymptotically always prefer to explain it as an additive outlier or as an innovative outlier. The following result describes how we can choose the hyper-parameters of the model to avoid this. The idea is to look at the particle filter weights for describing an extreme observation as either an additive or an innovative outlier, if that is possible, and ensuring they are of similar order to each other.

Theorem 4: To simplify notation we drop the temporal subscripts and let the prior for the hidden state $\mathbf{X}_t$ be $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and the observation at time $t+1$ be $\boldsymbol{Y}$. For either an additive anomaly for component i or an innovative anomaly for component j we can standardise the size of the anomaly to define δ such that

$$\mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu}=\frac{\delta\mathbf{e}_i}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\ \text{ or }\ \mathbf{Y}-\mathbf{C}\mathbf{A}\boldsymbol{\mu}=\frac{\delta\mathbf{C}^{(:,j)}}{\sqrt{\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}},$$



(a) $t=100$ (b) $t=101$ (c) Full data

Fig. 3: Robust particle filter output at various times. Additive anomalies are denoted by red points, innovative anomalies by blue lines. Grey observations are yet to be observed.

If the shape parameters for the prior for the precision of all anomalies are set to be the same, that is $a_1=...=a_p=b_1=...=b_q=c$, and if the prior mean for the precision of each anomaly is chosen to be

$$\tilde{\sigma_i}=\Sigma_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\ \text{ and }\ \hat{\sigma_j}=\Sigma_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)},$$

then, to leading order, the particle weights of additive and innovative anomalies are asymptotically proportional to

$$\frac{c^c\frac{1}{M}r_i\frac{\Gamma(c+\frac{1}{2})}{\Gamma(c)}\exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\delta^2}{2}\right)^c}\ \text{ and }\ \frac{c^c\frac{1}{M}s_j\frac{\Gamma(c+\frac{1}{2})}{\Gamma(c)}\exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\delta^2}{2}\right)^c}.$$

respectively, as $\delta\to\infty$.

The choice of hyper-parameters given in this theorem leads to all components being given equal asymptotic importance weight under an arbitrarily large anomaly. Setting all the a_is and b_js to the same constant is advisable due to the fact that the convolution of two t-distributions whose means drift further and further apart yields two stable, i.e. non-vanishing modes if and only if they have the same scale parameter.

While $\hat{\boldsymbol{\Sigma}}^{-1}$ is not fixed but time dependent, it nevertheless converges to a limit under an observable Kalman filter model. In practice, we therefore use this limit to set $\tilde{\sigma}_i$ and $\hat{\sigma}_j$.

C. Example 1 – revisited

The proposed filter can be applied to the data displayed in Figure 1a to detect anomalies in an online fashion. It is worth pointing out that the filter re-evaluates past anomalies as more data becomes available. This can be seen in Figure 3: When initially encountering the anomaly at time $t=100$ the filter gives approximately equal weight to the possibility of it being an additive outlier and to it being an innovative one. It is only when the next observation becomes available, that the filter (correctly) classifies it as an innovative anomaly. Note that only $N=20$ particles were used and only $M=1$ descendent of each anomaly type was sampled per particle.

IV. Particle Filter With Back-Sampling – CE-BASS

As mentioned in the introduction, it is possible that innovative outliers may not immediately be observed. One such example are innovative outliers in the trend component of the model described in (3). The filter as described in Algorithm 1 can not deal with such anomalies as it only inflates the variance of the innovative process at time t when there is

evidence from $\boldsymbol{Y}_t$ that an outlier occurred. We remedy this by back-sampling particles representing innovative outliers at a later time once we have more observations, and therefore more evidence for an anomaly is available. This can be done using nearly identical approximation strategies as used in the previous section and allows to relax the assumptions made in the previous section that $\mathbf{C}$ has no columns that just contain zeros, to only requiring that the system be observable.

A. Back-Sampling Particles Using the Last $k+1$ Observations

The proposed back-sampling strategy at time t consists of sampling particles for $(\mathbf{V}_{t+1-k}, ...\mathbf{V}_{t+1}, \mathbf{W}_{t+1-k}, ..., \mathbf{W}_{t+1})$ given a $N(\boldsymbol{\mu}_{t-k}, \boldsymbol{\Sigma}_{t-k})$ filtering distribution for $\mathbf{x}_{t-k}$ and observations $\mathbf{Y}_{t-k+1}, ..., \mathbf{Y}_{t-k}$. Specifically, we sample particles with an innovative single anomaly in $\mathbf{W}_{t+1-k}$ assuming no other innovative anomalies or additive anomalies. Conditional on these augmented particles classical Kalman updates can once more be used as shown in Algorithm 2.

At each iteration of Algorithm 2 we first simulate candidate weighted particles. At time t, for each particle at time $t-1$ we calculate the candidate particle that corresponds to no outlier at time t (Sample_typical in Step 4), and also simulate M candidate particles for each possible type of additive outlier (Sample_additive in Step 6). These can be carried out as before. Simulating candidate particles for innovative outliers is different and involves the idea of back-sampling. The algorithm has a user-defined maximum horizon ($max\_horizon$). For each $k = 1, \ldots, max\_horizon$ we consider all particles at time $t-k$ and simulate a set of descendants which have an innovative outlier at time $t-k+1$ and then no further outliers until time t. This is performed at step 14 with the Inn_Des function. This function outputs a set of M sample values for $\boldsymbol{V}$ and $\boldsymbol{W}$ at time $t-k+1$ for each type of innovative outlier, and each sample has an associated importance sampling weight. Importantly as we are comparing new particles proposed from old particles at different times in the past, the importance sampling weights need to include a factor that estimates the evidence, i.e. the marginal probability of the data, at the time of the old particle – see Section IV in the Supplementary Material for a more detailed explanation for this. We calculate the usual particle filter estimate of the evidence at each iteration in steps 21 to 24. The weights of these particles are down-weighted by a factor of $1/max\_horizon$ to account for the fact that the same innovative anomaly will be proposed multiple times.

After obtaining the full set of candidates, we resample them with probability proportional to their weight, for example using stratified resampling [14], and then use the Kalman Filter update to obtain the corresponding filtered mean and variance for the state at time t. For the back-sampled particles from time $t-k+1$ for $k > 1$ we need to apply the Kalman Filter update for k time steps, and this is done under the particle's assumption of no outliers at times $t-k+2, \ldots, t$.

Algorithm 1 is a special case of Algorithm 2 which arises from setting the maximum horizon to 1. The Sample_Particles function in Algorithm 1 corresponds to the simulation of candidates for no outlier, an additive outlier or an innovative outlier that are listed separately in Algorithm 2.

Algorithm 2 Particle Filter (With Back Sampling) – CE-BASS

Input: An initial state estimate $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.
A number of descendants, $M' = M(p+q)+1$.
A number of particles to be maintained, N.
A stream of observations $\mathbf{Y}_1, \mathbf{Y}_2, ...$
Initialise: Set *Particles*$(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, 1)\}$
$EV(t) = 1$
Set $max\_horizon$

1: **for** $t \in \mathbb{N}$ **do**
2: *Cand* $\leftarrow \{\}$
3: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in$ *Particles*(t) **do**
4: $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_typical$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_{t+1}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
5: *Cand* $\leftarrow$ *Cand* $\cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot EV(t), 1)\}$
6: $Add\_Des \leftarrow$ Sample_additive$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_{t+1}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$
7: **for** $(\mathbf{V}, \mathbf{W}, prob) \in Add\_Des$ **do**
8: *Cand* $\leftarrow$ *Cand* $\cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot EV(t), 1)\}$
9: **end for**
10: **end for**
11: **for** $k \in \{1, ..., max\_horizon\}$ **do**
12: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in$ *Particles*$(t-k+1)$ **do**
13: $\tilde{\mathbf{Y}} \leftarrow \left[\mathbf{Y}_{t-k+2}^T, ..., \mathbf{Y}_{t+1}^T\right]^T$
14: $Inn\_Des \leftarrow$ BS_inn$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M, k)$
15: **for** $(\mathbf{V}, \mathbf{W}, prob) \in Inn\_Des$ **do**
16: *Cand* $\leftarrow$ *Cand* $\cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, \frac{prob \cdot EV(t+1-k)}{max\_horizon}), k)\}$
17: **end for**
18: **end for**
19: **end for**
20: $EV(t+1) \leftarrow 0$ ▷ Calculate estimate of evidence at time $t+1$
21: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, k) \in$ *Cand* **do**
22: $EV(t+1) \leftarrow EV(t+1) + prob/|$*Cand*$|$
23: **end for**
24: *Descendants* $\leftarrow$ Resample$(N,$ *Cand*$)$ ▷ Resample particles
25: *Particles*$(t) \leftarrow \{\}$ ▷ Calculate $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\Sigma}_{t+1}$ for each particle
26: **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, k) \in$ *Descendants* **do**
27: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+2-k}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}\boldsymbol{\Sigma}_A, \mathbf{W}\boldsymbol{\Sigma}_I)$
28: **if** $k > 1$ **then**
29: **for** $i \in \{2, ..., k\}$ **do**
30: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+1+i-k}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
31: **end for**
32: **end if**
33: *Particles*$(t+1) \leftarrow$ *Particles*$(t+1) \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma})\}$
34: **end for**
35: **end for**

We now describe how we sample candidate particles which allow for innovative outliers in Step 14 of Algorithm 2. The idea is that we can use the same idea as previously, but for a larger state-space model that considers jointly all the observations since time $t-k+1$.

To sample a particle with an innovative anomaly in the jth component of $\mathbf{W}_{t+1-k}$, we define an augmented observation vector $\tilde{\mathbf{Y}}_{t+1-k}^{(k)} = (\mathbf{Y}_{t+1-k}^T, ..., \mathbf{Y}_{t+1}^T)^T$. This is normally distributed with mean $\tilde{\mathbf{C}}^{(k)}\mathbf{A}\boldsymbol{\mu}_{t-k}$ and variance

$$\tilde{\mathbf{C}}_0^{(k)}\mathbf{A}\boldsymbol{\Sigma}_{t-k}\mathbf{A}^T\left(\tilde{\mathbf{C}}_0^{(k)}\right)^T + \sum_{i=0}^{k}\left[\tilde{\mathbf{C}}_i^{(k)}\mathbf{V}_{t+1-k+i}^{-1}\boldsymbol{\Sigma}_A\left(\tilde{\mathbf{C}}_i^{(k)}\right)^T\right] + \tilde{\mathbf{R}}^{(k)},$$

where

$$\tilde{\mathbf{C}}_i^{(k)} = \mathbf{C}\left(\mathbf{0}_{q\times iq}, \left(\mathbf{A}^0\right)^T, ..., \left(\mathbf{A}^{k-i}\right)^T\right)^T \tag{6}$$

for $0 \leq i \leq k$ denote the augmented matrices mapping the hidden states and innovations to the observations and

$$\tilde{\mathbf{R}}^{(k)} = \begin{bmatrix} \mathbf{V}_{t+1-k}^{-1}\boldsymbol{\Sigma}_A & 0 & \ddots \\ 0 & \ddots & 0 \\ \ddots & 0 & \mathbf{V}_{t+1}^{-1}\boldsymbol{\Sigma}_A \end{bmatrix}$$

In a similar spirit, we define the augmented predictive variance $\hat{\mathbf{\Sigma}}^{(k)}$ to be

$$\tilde{\mathbf{C}}_0^{(k)}\mathbf{A}\mathbf{\Sigma}_{t-k}\mathbf{A}^T\left(\tilde{\mathbf{C}}_0^{(k)}\right)^T+\sum_{i=0}^{k}\left[\tilde{\mathbf{C}}_i^{(k)}\mathbf{\Sigma}_A\left(\tilde{\mathbf{C}}_i^{(k)}\right)^T\right]+\mathbf{I}_{k+1}\otimes\mathbf{\Sigma}_A.$$

As a result of this reformulation, we retrieve update equations consisting of a single Kalman step, albeit with slightly different dimensions of the observation, $(k+1)p$ instead of p. It is therefore possible to use the sampling procedure for innovative outliers introduced in Section III-A providing $\left(\tilde{\mathbf{C}}^{(k)}\right)^{(:,j)}\neq\mathbf{0}$. This consists of sampling particles for $\tilde{\mathbf{W}}_{t+1-k}^{(j,j)}$ from

$$\hat{\sigma_j}\Gamma\left(b_j+\frac{1}{2},b_j+\frac{\hat{\sigma_j}}{2\mathbf{\Sigma}_I^{(j,j)}}\left(\frac{\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\right)^{(j,:)}\left(\hat{\mathbf{\Sigma}}^{(k)}\right)^{-1}\tilde{\mathbf{z}}_{t+1-k}^{(k)}}{\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\mathbf{\Sigma}}^{(k)}\right)^{-1}\tilde{\mathbf{C}}^{(k)}\right)^{(j,j)}}\right)^2\right).$$

for the residual $\tilde{\mathbf{z}}_{t+1-k}^{(k)}\tilde{\mathbf{Y}}_{t+1-k}^{(k)}-\tilde{\mathbf{C}}^{(k)}\mathbf{A}\boldsymbol{\mu}_{t-k}$. The associated weight is given in Theorem 5 in the Supplementary Material. For details of how we choose the hyper-parameters for this proposal see Section I in the Supplementary Material.

A range of observations guide the choice of the maximum horizon. We assume that the Kalman model is observable, i.e. that there exists a k such that the matrix $\left[(\mathbf{C})^T,(\mathbf{CA})^T,...,\left(\mathbf{CA}^k\right)^T\right]$ has full column rank. Let k^* denote the lowest such k. We suggest choosing the maximum horizon so that it is at least equal to or bigger than k^*, as any innovative anomaly capable of influencing the observations must do so within k^* time steps. Increasing the maximum horizon further can be beneficial, as it allows detection of weaker, but persistent, innovative anomalies (e.g. weak changes in mean). However, this comes at an increased computational cost. It can therefore be recommended to set it to as large a value as is computationally feasible.

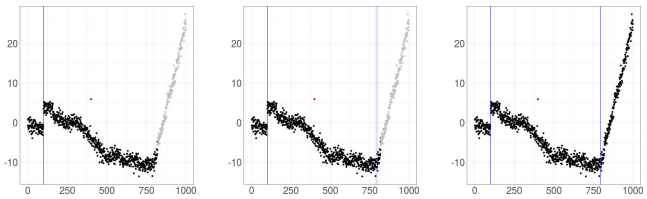
In some situations we may wish to only consider back-sampling a set of previous time-points. This could be to reduce computational cost, or to account for the above proposal distribution not being well defined for k if $\left(\tilde{\mathbf{C}}^{(k)}\right)^{(:,j)}$ is the $\mathbf{0}$ vector. This is possible if we change step 11 of Algorithm 2 to consider only k within some subset of $\{1,\ldots,max\_horizon\}$.

B. Example

With back-sampling, we are now able to tackle the example from Figure 1b. We used a maximum horizon of 40 We maintained $N=40$ particles and sampled $M=1$ descendants of each type. The output of the particle filter can be seen in Figure 4. As before, the filter updates its output as new observations become available. Whilst the trend innovation occurs at time $t=800$, the anomaly is first detected around time $t=820$. Even then, there is a large amount of uncertainty regarding the precise location of the anomaly which only gets resolved at a later time.

C. Computational Cost

First and foremost, CE-BASS is fully on-line; i.e. its computational cost does not increase in time. This constant computational cost of each individual step is $O(NM(p^3+q^3))$ when no back-sampling is used and dominated by the cost of matrix multiplications/inversions. Back-sampling at a horizon k increases the dimension of the predictive variance matrix $\hat{\mathbf{\Sigma}}$ from $p\times p$ to $pk\times pk$. Since it has to be inverted for $k=1,...,max\_horizon$, the computational cost is proportional to $max\_horizon^4$.



(a) $t=820$ (b) $t=821$ (c) Full data

Fig. 4: Robust particle filter output at various times. Additive anomalies are denoted by red points, innovative anomalies by blue lines. Grey observations are yet to be observed.

When processing Example 1 on a standard laptop, our C++ implementations took an average of $3.3ms$ (CE-BASS), $0.7ms$ (IORKF from [9]), $0.8ms$ (AORKF from [9]), $1.1ms$ [8], and $0.7ms$ (classical Kalman filter) for each iteration. This increased to $51.8ms$ (CE-BASS), $0.8ms$ (IORKF from [9]), $0.8ms$ (AORKF from [9]), $1.2ms$ [8], and $0.8ms$ (classical Kalman filter) per iteration for the second example.

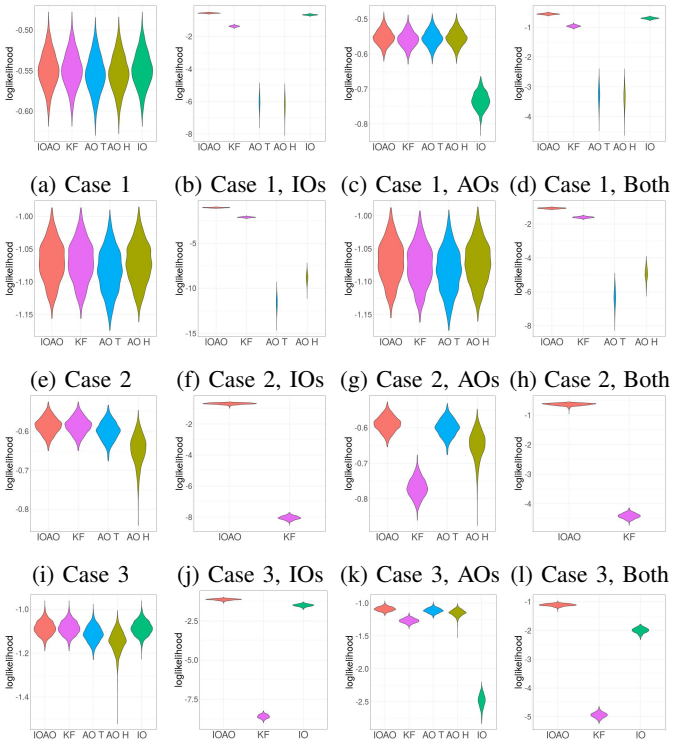
V. Simulations

We now turn to comparing CE-BASS against other methods. In particular, we compare against the t-distribution based additive outlier robust filter by [8], the Huberisation, i.e. truncation, based additive outlier robust filter by [9], the Huberisation based innovative outlier robust filter by [9], and the classical Kalman Filter [4]. All these algorithms are implemented in the accompanying package.

We consider four different models and generate 1000 observations for each. For each of the four models, we consider a case in which no anomalies are present, a case in which only additive anomalies are present, a case in which only innovative anomalies are present, and a case in which both additive and innovative anomalies are present. When anomalies are added, they are added at times $t=100$, $t=300$, $t=600$, and $t=900$. Specifically we considered the following three models:

1) The model of Example 1 with $\sigma_A=1$ and $\sigma_I=0.1$. We consider a case with only additive outliers, a case with only innovative outliers, and a case where an additive outlier at $t=100$, is followed by two innovative outliers at times $t=300$ and $t=600$, which were then followed by an additive outlier at time $t=900$. To simulate additive anomalies, we set $V_t^{\frac{1}{2}}\sigma_A\epsilon_t=10$ and to simulate the innovative outliers we set $W_t^{\frac{1}{2}}\sigma_I\nu_t=10$.
2) The random walk model with two measurements

$$Y_t^{(1)}=X_t+\left(V_t^{(1)}\right)^{\frac{1}{2}}\sigma_A^{(1)}\epsilon_t^{(1)},\qquad X_t=X_{t-1}+W_t^{\frac{1}{2}}\sigma_I\nu_t$$
$$Y_t^{(2)}=X_t+\left(V_t^{(2)}\right)^{\frac{1}{2}}\sigma_A^{(2)}\epsilon_t^{(2)},$$



(a) Case 1 (b) Case 1, IOs (c) Case 1, AOs (d) Case 1, Both

(e) Case 2 (f) Case 2, IOs (g) Case 2, AOs (h) Case 2, Both

(i) Case 3 (j) Case 3, IOs (k) Case 3, AOs (l) Case 3, Both

(m) Case 4 (n) Case 4, IOs (o) Case 4, AOs (p) Case 4, Both

Fig. 5: Violin plots for the average predictive log-likelihood of the five filters (IOAO: CE-BASS, KF: The classical Kalman Filter, AO T: [8], AO H: [9], IO: [9]) over the four different scenarios under a range of models. Higher values correspond to better performance. Methods are omitted on the graphs if they can not be applied to the setting or if their performance is too poor.

where $\sigma_A^{(1)} = \sigma_A^{(2)} = 1$ for $i = 1, 2$ and $\sigma_I = 0.1$. We consider a case with only additive outliers (one in the first component, then two in the second, then one in the first), a case with only innovative outliers, and a case where an additive outlier in the first component at time $t = 100$ is followed by two innovative outliers at times $t = 300$ and $t = 600$, which are then followed by an additive outlier in the second component at time $t = 900$. For additive anomalies, $(V_t^{(1)})^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)} = 10$ or $(V_t^{(2)})^{\frac{1}{2}} \sigma_A^{(2)} \epsilon_t^{(2)} = 10$ and for innovative outliers, $W_t^{\frac{1}{2}} \sigma_I \nu_t = 10$.

3) The model of Example 2 with $\sigma_A = 1$, $\sigma_I^{(1)} = 0.1$ and $\sigma_I^{(2)} = 0.01$. We consider a case with only additive outliers, a case with only innovative outliers (one in the second component, then one in the first, then one in the second, then one in the first), and a case with an additive outlier at $t = 100$, followed by an innovative outlier affecting the first component of the hidden state at times $t = 300$, followed by an innovative outlier affecting the second component of the hidden state at times $t = 600$, followed by an additive outlier at time $t = 900$. The additive anomalies were instances where $V_t^{\frac{1}{2}} \epsilon_t = 30$ and the innovative outliers were instances where $(W_t^{(1)})^{\frac{1}{2}} \eta_t^{(1)} = 100$ or $(W_t^{(2)})^{\frac{1}{2}} \eta_t^{(2)} = 500$.

4) An extension of Example 2 where the position is also observed. The equations governing the hidden state are as before whilst the equations governing the observations are

$$Y_t^{(1)} = X_t^{(1)} + \left(V_t^{(1)}\right)^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)},$$
$$Y_t^{(2)} = X_t^{(2)} + \left(V_t^{(2)}\right)^{\frac{1}{2}} \sigma_A^{(2)} \epsilon_t^{(2)},$$

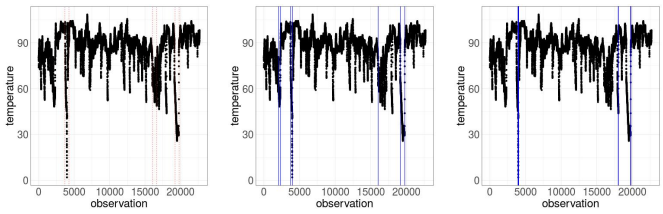
where $\sigma_A^{(1)} = \sigma_A^{(2)} = 1$. We consider a case with only additive outliers (in the first component only), a case with only innovative outliers (one in the second component, then one in the first, then one in the second, then one in the first), and a case with an additive outlier at time $t = 100$, followed by an innovative outlier affecting the first component of the hidden state at time $t = 300$, followed by an innovative outlier affecting the second component of the hidden state at time $t = 600$, followed by an additive outlier at time $t = 900$. For additive anomalies, $(V_t^{(1)})^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)} = 30$ and for innovative outliers, $(W_t^{(1)})^{\frac{1}{2}} \sigma_I^{(1)} \eta_t^{(1)} = 100$ or $(W_t^{(2)})^{\frac{1}{2}} \sigma_I^{(2)} \eta_t^{(2)} = 500$.

We evaluate the different methods based on average predictive log-likelihood and average predictive mean squared error. That is we calculate the one step-ahead predictive distribution for the next observation, and respectively evaluate the log-predictive density of the observation or evaluate the square error of the mean of the predictive distribution, and then average these quantities over the observations. We exclude all observations corresponding to anomalies from the calculation of these averages since the filters can not be expected to predict them. When calculating the average mean squared error we additionally remove one observation after the anomaly in the first setting and two observations in the third setting from the performance metric. This is to give the filter enough information to determine which type of anomaly the outlier corresponds to and return to a unimodal posterior: the MSE is a less informative metric for multimodal posteriors, as it is minimised at the posterior mean and this can be in a region of negligible posterior mass.

The average log-likelihoods across all models can be found in Figure 5, while the qualitatively very similar results for the mean squared error can be found in the Supplementary Material. We see that the performance of CE-BASS compares favourably with that of the competing methods. In particular it is as accurate as the Kalman filter in the absence of anomalies and is more accurate than the additive outlier and innovative outlier robust filters even when only additive or innovative outliers are present, i.e. the settings for which these algorithms were designed.

VI. Application

We now apply CE-BASS with two different types of model to illustrate how CE-BASS can be used on real datasets. The first dataset is a labelled benchmark dataset which consists of temperature readings on a large industrial machine. Here, we will use a model which considerably restricts the movements of the hidden states when no anomalies are present, and thus



(a) Data + labels (b) CE-BASS output (c) IO-RKF output

Fig. 6: Machine temperature dataset. The known anomalous regions are shown by the red regions. The estimated locations of innovative outliers are shown by blue vertical lines. For the IO-RKF some of these are at very similar times – and these can only be noticed by eye through the wider vertical lines on the plot.

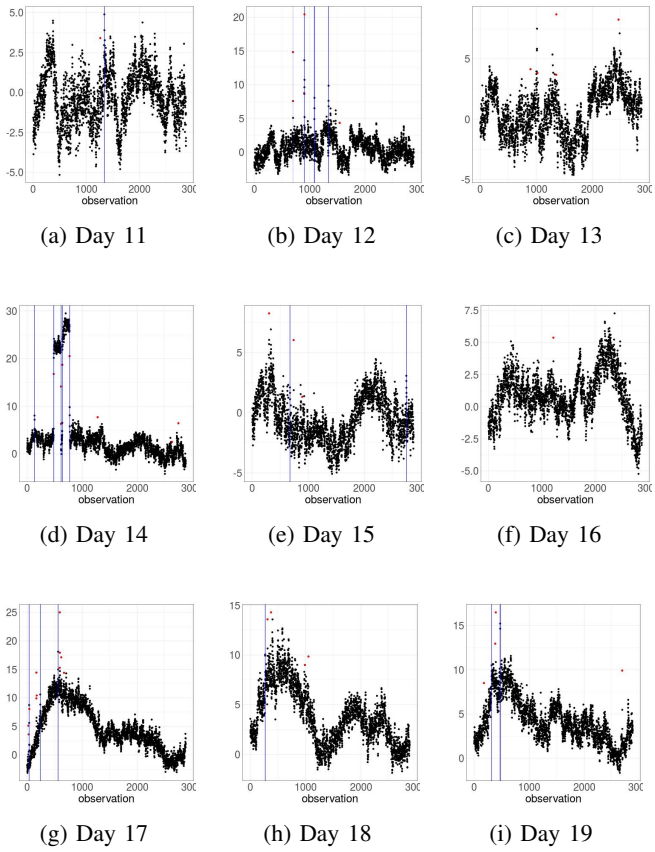
emulates a changepoint model. The second is an unlabelled dataset which consists of repeated throughput measurements on a router. For that application we will use a model which has a considerable amount of flexibility and where the hidden states tend to follow the observations and therefore detect localised anomalies.

A. Machine Temperature Data

We now apply CE-BASS to the machine temperature data taken from the Numenta Anomaly Benchmark (NAB, [23]) which can be accessed at *https://github.com/numenta/NAB*. The data consists of over 20000 readings from a temperature sensor on a large industrial machine and is displayed in Figure 6a along the three periods of anomalous behaviour labelled by an engineer. The first corresponds to a planned shutdown and the second to an early warning sign of the third anomaly – a catastrophic failure.

In order to do so, we use the random walk model from Example 1 with the aim of detecting persistent changes in mean. We therefore use a maximum backsampling horizon of 250 but to reduce computational cost we only consider back-sampling at a sub-set of earlier times, so in Step 11 of Algorithm 2 we consider only $k \in \{1, 5, 10, 20, 40, 80, 150, 250\}$ and fix $\sigma_I = 1/10000\sigma_A$ to ensure that long and weak anomalies will not be interpreted as a persistent shift in the typical state. We use the first 15% of the data, marked by [23] as training data, to estimate the standard deviation σ_A as well as the initial mean μ_0 using the median absolute deviation and the median respectively. Using robust covariance methods we also detect very strong auto-correlation ($\rho = 0.99$) and therefore took the default probabilities for anomalies to the power of $\frac{1}{1-\rho}$.

The results of this analysis can be seen in Figure 6b. We note that all anomalies flagged by the engineer are also being detected by CE-BASS. Two additional innovative anomalies around a prolonged drop which preceded the planned shutdown are also detected. They could be a false positive or an early warning sign of an anomaly prevented by the shutdown which has not been noticed by the engineer. For comparison, five anomaly detection methods were tested on this data in



(a) Day 11 (b) Day 12 (c) Day 13

(d) Day 14 (e) Day 15 (f) Day 16

(g) Day 17 (h) Day 18 (i) Day 19

Fig. 7: CE-BASS applied to 9 days of de-seasonalised router data. Lines correspond to innovative anomalies, i.e. spikes or level shifts.

[23]: only one of these was able to detect all three anomalies and that method had three additional false positives.

We also applied the innovative outlier robust Kalman filter by [9] using the same values of σ_A, σ_I, A, C, and Σ_0. The initial mean μ_0 was set to the value of the first observation. For the purpose of comparison, we chose the threshold for an anomaly such that the number of detected anomalies is equal to that found by CE-BASS. The result of this analysis is displayed in Figure 6c. The detected anomalies overlap with just the first and last of the known anomalies, and it picks up an additional anomaly between the second and third known anomalies. It can be seen that the detected anomalies correspond to the largest jumps in observed values. This highlights the practical value of CE-BASS's back-sampling as it is able to detect weaker, but persistent changes, such as the second anomalous window of the machine temperature dataset.

B. Router Data

The online analysis of aggregated traffic data on servers is an important challenge in both predictive maintenance and cyber security. This is because anomalies in throughput can point towards problems in the network such as malfunctions or malicious behaviour. Detecting anomalies as soon as possible

therefore means that the root cause can be addressed more quickly – potentially even before user experience is affected or harm caused.

In this section, we consider 19 days worth of (unlabelled) data that represents the input data rate observed at an IP router interface in the core of a telecommunications network. The data is gathered at a frequency of one observation every 30 seconds. To preserve confidentiality, we de-seasonalised the data for days 11 to 19 using a seasonality model trained on days 1 to 10 and, for the purpose of this paper, consider only the de-seasonalised data for days 11 to 19 which can be found in Figures 7a to 7i. The main features apparent in the daily series are spikes, outliers, and changepoints. In order to capture these, we use an AR(1) model with slowly changing mean to model the observations Y_t. Formally, we used the model

$$\begin{aligned} Y_t &= X_t^{(1)} + X_t^{(2)} + V_t \sigma_A \epsilon_t, \\ X_t^{(1)} &= X_{t-1}^{(1)} + W_t^{(1)} \sigma_I^{(1)} \eta_t^{(1)}, \\ X_t^{(2)} &= \rho X_{t-1}^{(2)} + W_t^{(2)} \sigma_I^{(2)} \eta_t^{(2)}. \end{aligned}$$

Here, anomalies in ϵ_t correspond to isolated outliers, anomalies in $\eta_t^{(1)}$ correspond to level shifts and outliers in $\eta_t^{(2)}$ correspond to spikes.

We use the first 1000 observations of the first day, to estimate the hyper-parameters. We first used robust loess-smoothing to obtain a smoothed signal $\hat{y}_t$ from the original time series y_t. Taking a robust estimate of the variance of $\hat{y}_t - \hat{y}_{t-1}$, we estimated $\sigma_I^{(1)} = 0.0157$. Using a robust AR(1) regression on the residuals $y_t - \hat{y}_t$, we further estimated $\sigma_I^{(2)} = 0.516$ and $\rho = 0.815$. We then set $\sigma_A = 1/10\sigma_I^{(1)} = 0.0516$,

The result obtained from running CE-BASS with these parameters on the daily router data is displayed in Figures 7a to 7i. A large number of anomalies are flagged, including a large number of outliers and spikes, but also some level shifts (Day 14). Discussion with engineers highlighted that the anomalies detected matched well with their knowledge of the data. This shows CE-BASS's ability to return a large number of diverse features which can be used as inputs to a supervised algorithm should labels become available.

VII. Discussion

We have presented CE-BASS, a robust particle filter algorithm that can deal with both innovative and additive outliers. The main limitation of this algorithm is that it assumes only a single outlier, that is one affecting a single component of either the additive or innovative noise, is possible at any time-step. This assumption is needed to obtain our efficient proposal distribution for particles, and one important extension of our work would be to relax this assumption. We show in simulations in the Supplementary Material that that the performance of CE-BASS can deteriorate in situations where multiple outliers occur simultaneously. Particular care should be taken if we have transformed the model or observation equation to remove correlations in the noise, as described in Section II. In this case an outlier in, say, one component of the observation vector could appear as an outlier affecting multiple components of the transformed observation.

VIII. Acknowledgements

This work was supported by EPSRC grant numbers EP/N031938/1 (StatScale) and EP/L015692/1 (STOR-i). The authors also acknowledge British Telecommunications plc (BT) for financial support, David Yearling and Trevor Burbridge in BT Research for discussions and Gaetano Romano for help with the machine temperature data.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[2] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[3] A. T. M. Fisch, I. A. Eckley, and P. Fearnhead, "A linear time method for the detection of point and collective anomalies," *arXiv preprint arXiv:1806.01947*, 2018.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[5] M. A. Gandhi and L. Mili, "Robust Kalman filter based on a generalized maximum-likelihood-type estimator," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2509–2520, 2009.

[6] Y. Huang, Y. Zhang, N. Li, Z. Wu, and J. A. Chambers, "A novel robust Student's t-based Kalman filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 3, pp. 1545–1554, 2017.

[7] J.-A. Ting, E. Theodorou, and S. Schaal, "Learning an outlier-robust Kalman filter," in *European Conference on Machine Learning*. Springer, 2007, pp. 748–756.

[8] G. Agamennoni, J. I. Nieto, and E. M. Nebot, "An outlier-robust Kalman filter," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 1551–1558.

[9] P. Ruckdeschel, B. Spangl, and D. Pupashenko, "Robust Kalman tracking and smoothing with propagating and non-propagating outliers," *Statistical Papers*, vol. 55, no. 1, pp. 93–123, 2014.

[10] G. Chang, "Robust Kalman filtering based on Mahalanobis distance as outlier judging criterion," *Journal of Geodesy*, vol. 88, no. 4, pp. 391–401, 2014.

[11] Y. Huang, Y. Zhang, Y. Zhao, and J. A. Chambers, "A novel robust Gaussian-Student's t mixture distribution based Kalman filter," *IEEE Transactions on Signal Processing*, 2019.

[12] G. Kitagawa, "Non-Gaussian state—space modeling of nonstationary time series," *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1032–1041, 1987.

[13] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *IEE proceedings F (Radar and Signal Processing)*, vol. 140, no. 2. IET, 1993, pp. 107–113.

[14] P. Fearnhead and H. R. Künsch, "Particle filters and data assimilation," *Annual Review of Statistics and Its Application*, vol. 5, no. 1, pp. 421–449, 2018.

[15] R. Chen and J. S. Liu, "Mixture Kalman filters," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 3, pp. 493–508, 2000.

[16] M. K. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filters," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 590–599, 1999.

[17] A. Fisch, D. Grose, I. Eckley, P. Fearnhead, and L. Bardwell, *RobKF: Innovative and/or Additive Outlier Robust Kalman Filtering*, 2021. [Online]. Available: https://CRAN.R-project.org/package=RobKF

[18] P. Fearnhead and G. Rigaill, "Changepoint detection in the presence of outliers," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 169–183, 2019.

[19] H. Maeng and P. Fryzlewicz, "Detecting linear trend changes and point anomalies in data sequences," *arXiv preprint arXiv:1906.01939*, 2019.

[20] P. Fearnhead and P. Clifford, "On-line inference for hidden Markov models via particle filters," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 4, pp. 887–899, 2003.

[21] R. Douc and O. Cappé, "Comparison of resampling schemes for particle filtering," in *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.* IEEE, 2005, pp. 64–69.

[22] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[23] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 38–44.