INNOVATIVE AND ADDITIVE OUTLIER ROBUST KALMAN FILTERING WITH A ROBUST PARTICLE FILTER

A PREPRINT

Alexander T. M. Fisch
Lancaster University
Lancaster, United Kingdom

Idris A. Eckley
Lancaster University
Lancaster, United Kingdom

Paul Fearnhead
Lancaster University
Lancaster, United Kingdom

July 8, 2020

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.

Keywords Kalman Filter · Anomaly Detection · Particle Filtering · Robust Filtering

1 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 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 of outliers are the 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 residuals, and [10] 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: if the outlying observation was caused by an additive anomaly, the state will be close to the prior, whereas if it was caused by an innovative anomaly, the state would be far from it.

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 [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 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.

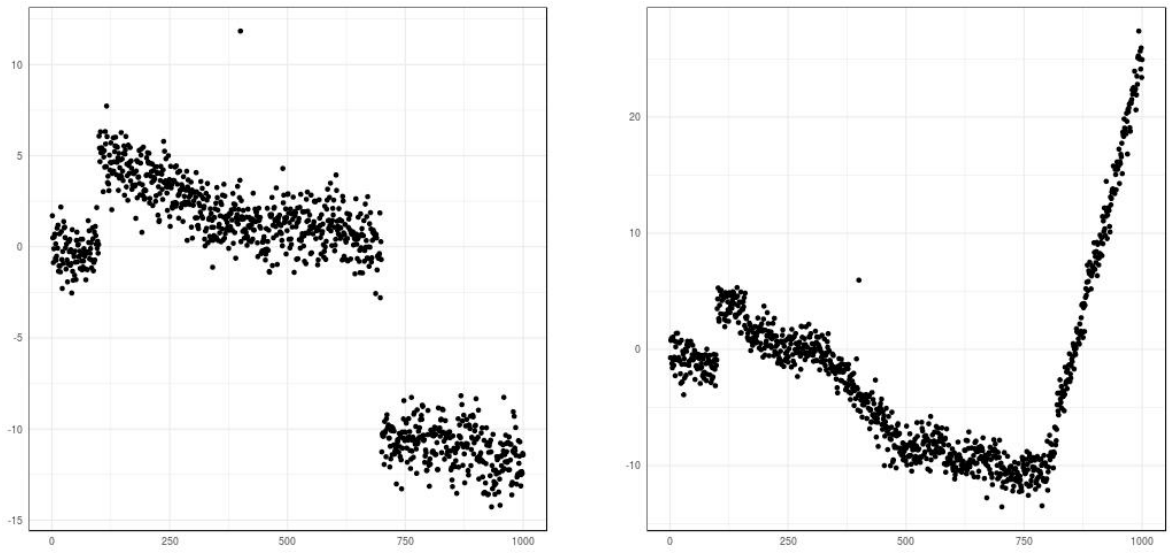
Another important challenge addressed by this paper is that certain innovative outliers can not immediately be detected. An innovative outlier in a latent trend component for instance can cause a trend changes which may only become apparent – i.e. produce a visible outlier in the observations – many 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 2. The model is shown to be very similar to that considered by [11]. We then introduce the proposal distribution for the scale of the noise in Section 3, before extending it to anomalies which are not immediately identifiable in Section 4. The proposed filter is compared to others in Section 5 and applied to router data and a benchmark machine temperature data-set in Section 6. 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 from `https://github.com/Fisch-Alex/Robkf`. Derivations of theoretical results and complete pseudocode are available in the appendix.

2 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

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



(a) Random walk

(b) Random walk with trends

Figure 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 diagonal matrices $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_I$ denote the covariance of the additive and innovation noise respectively. 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 changepoints and outliers, similar to the problem considered by [17]. 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.

Example 2: A time series with changes in trend, level shifts, as well as outliers, similar to the model considered by [18]. It can be formulated as

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

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 4. A wide rage of other commonly used time series features, such as auto-correlation, moving averages, etc. can be incorporated in the model.

To infer the locations of anomalies we use the model

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

for $1 \leq i \leq p$ and $1 \leq j \leq q$. The 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 (if present) is given by the random variables $\tilde{\mathbf{V}}_t^{(i,i)} \sim \tilde{\sigma}_i \Gamma(a_i, a_i)$ 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.

3 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, 19] is summarised by the pseudocode in Algorithm 1.

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 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 growing the number of particles exponentially with t, Step 7 resamples the candidates to keep just N particles. The filtering distribution for each of these particles is then calculated using the Kalman Filter updates in step 10.

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: $\quad Candidates \leftarrow \{\}$
3: $\quad$ **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in Particles(t-1)$ **do**
4: $\quad\quad (\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_Particles$(M', \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
5: $\quad\quad Candidates \leftarrow Candidates \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob)\}$
6: $\quad$ **end for**
7: $\quad Descendants \leftarrow$ Subsample$(N, Candidates)$
8: $\quad Particles(t) \leftarrow \{\}$
9: $\quad$ **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob) \in Descendants$ **do**
10: $\quad\quad (\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new}) \leftarrow$ KF_Upd$(\mathbf{Y}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$
11: $\quad\quad Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})\}$
12: $\quad$ **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 [20], 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

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

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, 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 [19]. 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 $\hat{\mathbf{\Sigma}} = \mathbf{C}\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\mathbf{C}^T + \mathbf{\Sigma}_A + \mathbf{C}\mathbf{\Sigma}_I\mathbf{C}^T$ and use it throughout the remainder of this paper. We also begin by assuming that $\mathbf{C}$ contains no 0-columns. The proposal introduced in the following subsection also forms the basis of back-sampling introduced in Section 4, which allows to relax this on $\mathbf{C}$.

3.1 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

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

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,

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

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,

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

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

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

up to $O\left(\left(\tilde{\mathbf{V}}^{(i,i)}\right)^2\right)$. 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 the proposal distribution

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

for $\tilde{\mathbf{V}}^{(i,i)}$. More detailed derivations, including the associated weight are given by Theorem 1 in the appendix. This proposal has the property that as the observed anomaly in the ith component becomes larger, i.e. as

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

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

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

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

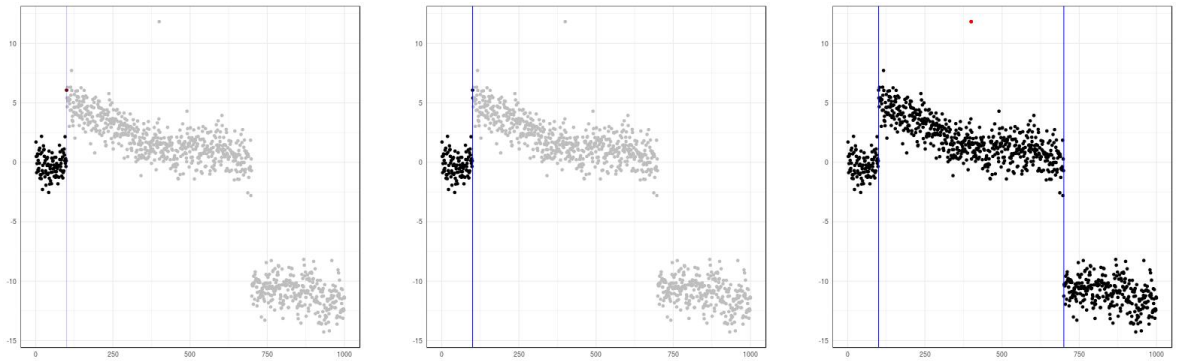
$$\tilde{\mathbf{W}}^{(j,j)} \sim \hat{\sigma_j}\Gamma\left(b_j + \frac{1}{2}, b_j + \frac{\hat{\sigma_i}}{2\mathbf{\Sigma}_I^{(j,j)}}\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)}\hat{\mathbf{\Sigma}}^{-1}(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu})}{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2\right)$$

and is of very similar form to the proposal distribution for particles with an additive outlier and well defined if $\mathbf{C}$ has no $\mathbf{0}$-columns. Further details, including the associated weight, are given in Theorem 2 in the appendix. 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 appendix.

3.2 Choices of Parameters

The choice of hyper-parameters, particularly $\hat{\sigma_i}$ and $\tilde{\sigma_i}$, has a significant effect of 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. The following result describes such a choice:



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

Figure 2: 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.

Theorem 4 *Let the prior for the hidden state $\boldsymbol{X}_t$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and an observation $\boldsymbol{Y}_{t+1} := \boldsymbol{Y}$ be available. When*

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

and $a_1 = ... = a_p = b_1 = ... = b_q = c$, the weights of additive and innovative anomalies are asymptotically proportional to

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

when

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

respectively, as $\delta \to \infty$

The above choice of hyper-parameters therefore leads to all components being given equal asymptotic importance weight under an anomaly they are able to account for. I.e. one which satisfies $\frac{\mathbf{C}^{(:,j)}}{\sqrt{\left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)}}} \delta = \mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu} = \frac{\delta \mathbf{e}_i}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}$. 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 $\hat{\sigma}_j$.

3.3 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 2: 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.

4 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 in the observation at the same time t that an outlier occurred. This can be remedied by back-sampling

particles representing innovative outliers at a later time, $t+k$, once more observations and therefore evidence for an anomaly are 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 from $\mathbf{C}$ not having any $\mathbf{0}$-columns to requiring that the system be observable.

4.1 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 a 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. It should be noted that Algorithm 1 is a special case of Algorithm 2 which arises from setting $\mathcal{B}_1 = ... = \mathcal{B}_q = \{1\}$.

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 $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, 1)\}$
Set $max\_horizon = \max(\cup_{i=1}^{q} \mathcal{B}_i)$

for $t \in \mathbb{N}^+$ **do**
  $Cand \leftarrow \{\}$ ▷ To Store Candidates
  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-1)$ **do**
    $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_typical$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
    $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$
    $Add\_Des \leftarrow$ Sample_additive$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$
    for $(\mathbf{V}, \mathbf{W}, prob) \in Add\_Des$ **do**
      $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$
    end for
  end for
  for $hor \in \{1, ..., max\_horizon\}$ **do**
    for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-hor)$ **do**
      $\tilde{\mathbf{Y}} \leftarrow \left[\mathbf{Y}_{t-hor+1}^T, ..., \mathbf{Y}_t^T\right]^T$
      $Inn\_Des \leftarrow$ BS_inn$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M, hor)$
      for $(\mathbf{V}, \mathbf{W}, prob) \in Inn\_Des$ **do**
        $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, hor)\}$
      end for
    end for
  end for
  $Desc \leftarrow$ Subsample$(N, Cand)$ ▷ Sampling proportional to $prob$
  $Particles(t) \leftarrow \{\}$
  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, hor) \in Descendants$ **do**
    $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+1-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$
    if $hor > 1$ **then**
      for $i \in \{2, ..., hor\}$ **do**
        $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+i-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
      end for
    end if
    $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob \cdot \frac{|Cand|}{|Desc|})\}$
  end for
end for

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

$$\tilde{\mathbf{C}}^{(k)}\left(\mathbf{A}\boldsymbol{\Sigma}_{t-k}\mathbf{A}^T + \tilde{\mathbf{Q}}^{(k)}\right)\left(\tilde{\mathbf{C}}^{(k)}\right)^T + \tilde{\mathbf{R}}^{(k)},$$

where $\tilde{\mathbf{C}}^{(k)} = \mathbf{C}\left(\left(\mathbf{A}^0\right)^T, ..., \left(\mathbf{A}^k\right)^T\right)^T$ denotes the augmented matrix mapping the hidden states to the observations,

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

and

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

In a similar spirit, we define the augmented predictive variance to be

$$\hat{\boldsymbol{\Sigma}}^{(k)} = \tilde{\mathbf{C}}^{(k)}\left(\mathbf{A}\boldsymbol{\Sigma}_{t-k}\mathbf{A}^T + \mathbf{I}_{k+1}\otimes\boldsymbol{\Sigma}_I\right)\left(\tilde{\mathbf{C}}^{(k)}\right)^T + \mathbf{I}_{k+1}\otimes\boldsymbol{\Sigma}_A.$$

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 3.1. This consists of sampling particles for $\tilde{\mathbf{W}}_{t+1-k}^{(j,j)}$ from

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

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 appendix.

As in Section 3.2, we want to give different particles equal weights if they explain anomalies equally well. In particular, we therefore want to balance out the weights given to the back-sampled particles and the descendants of particles with an anomaly sampled at time $t-k+1$ using just $\mathbf{Y}_{t+1-k}$. In order to do so, consider observations $\mathbf{Y}_{t+1}, ..., \mathbf{Y}_{t+1-k}$ which are such that they perfectly fit an innovative outlier in the ith innovative component at time $t-k+1$, i.e.

$$\tilde{\mathbf{Y}}_{t+1-k}^{(k)} - \left(\tilde{\mathbf{C}}^{(k)}\right)\mathbf{A}\boldsymbol{\mu}_{t-k} = \frac{\left(\tilde{\mathbf{C}}^{(k)}\right)^{(:,j)}}{\sqrt{\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}}}\delta.$$

As δ grows, the importance weight behaves as

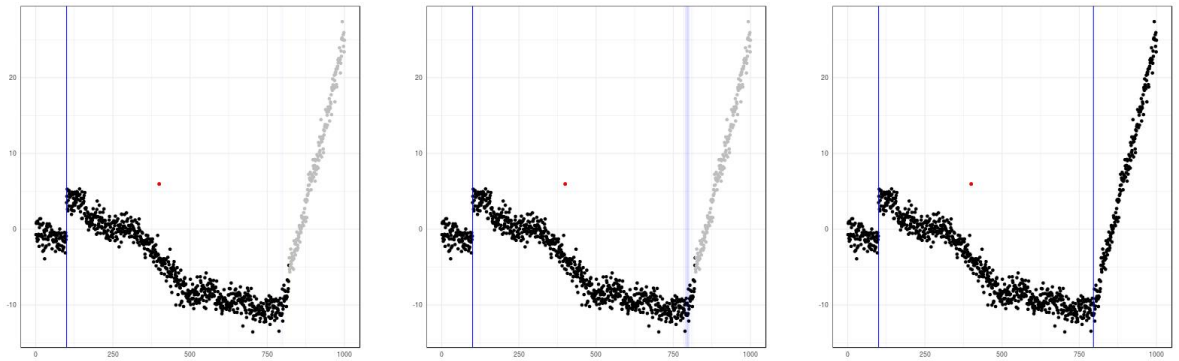
$$\frac{b_j^{b_j}\frac{1}{M}s_j\frac{\Gamma(b_j+\frac{1}{2})}{\Gamma(b_j)}\exp\left(-\delta^2\right)}{\left(\frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}}\delta^2\right)^{b_j}},$$

up to the likelihood term and the $\left(1-\sum_{i=1}^p r_i - \sum_{j=1}^q s_j\right)^k$ factor. However, these terms are also present in the weights of the descendants of the particles sampled at $t+1-k$ if no further anomaly was sampled at times $t+2-k, ..., t+1$. Therefore, setting

$$\hat{\sigma}_j = \boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}$$

results in the same asymptotic probabilities as the one obtained in Section 3.2. Given $\hat{\sigma}_j$ can only take a single value we set

$$\hat{\sigma}_j = \max_{k\in\mathcal{B}_j}\left(\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\mathbf{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\mathbf{C}}^{(k)}\right)\right)^{(j,j)}\right),$$



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

Figure 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.

where $\mathcal{B}_j \subset \mathbb{N}$ denotes the set of horizons used to back-sample the jth component of the $\mathbf{W}_t$.

A range of observations guide the choice of the sets $\mathcal{B}_j$ for $1 \leq j \leq q$. 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. It is advisable to choose the set $\mathcal{B}_j$ such that it contains at least one element greater or equal to k^*. The reason for this being that any innovative anomaly capable of eventually influencing the observations must do so within k^* observations from occurring. It should also be noted that a horizon h can only be in the set $\mathcal{B}_j$ if the jth column of the augmented mapping from the hidden states to the observations, $\tilde{\mathbf{C}}^{(h)}$, is non-zero as this is required by the proposal. Consequently, setting $\mathcal{B}_j = \left\{k \in \{1, ..., k^*\} : \left(\tilde{\mathbf{C}}^{(k)}\right)^{(:,j)} \neq \mathbf{0}\right\}$ is a natural choice.

4.2 Example

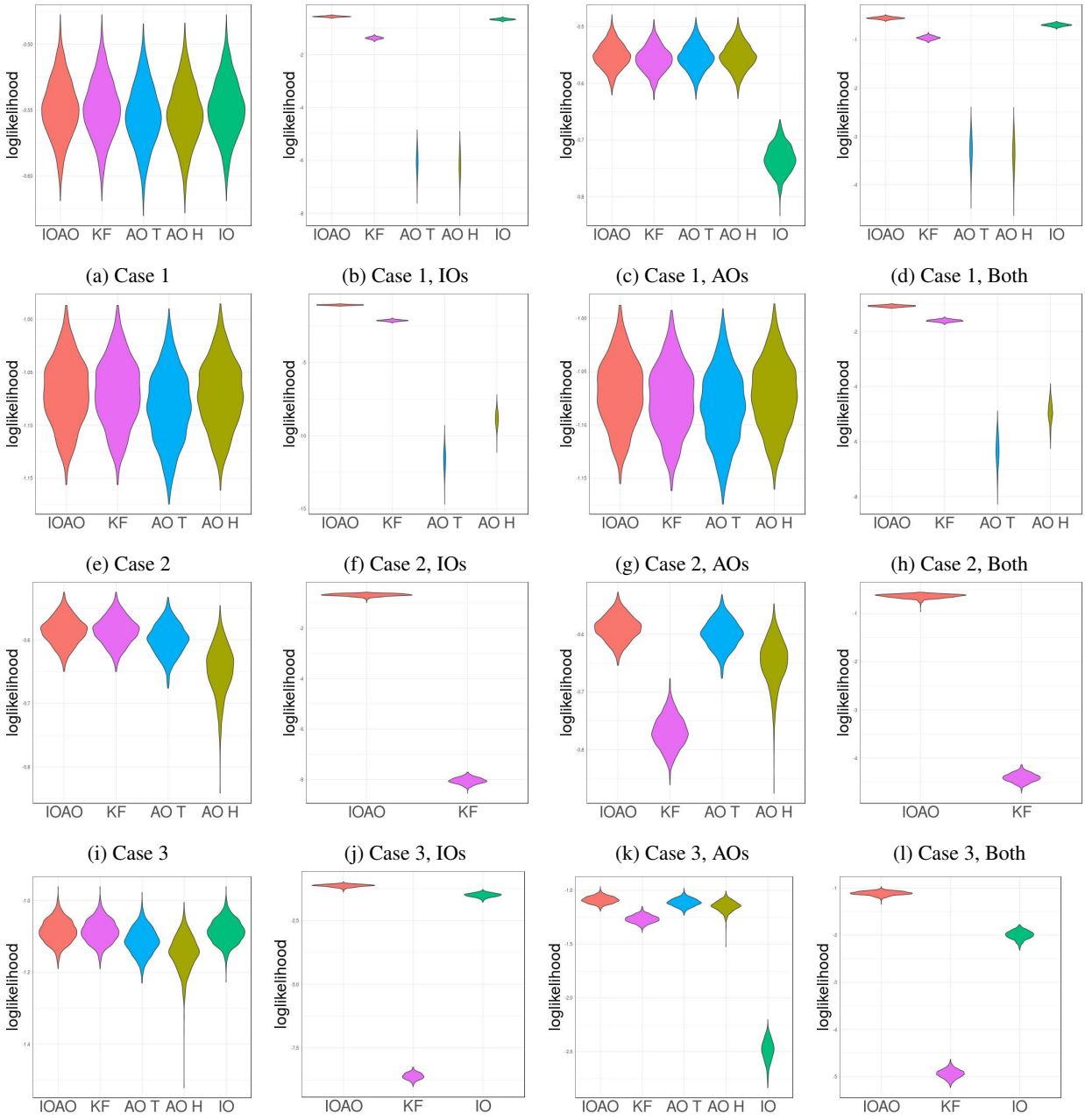
With back-sampling, we are now able to tackle the example from Figure 1b. We used $\mathcal{B}_1 = \{1, ..., 40\}$, $\mathcal{B}_2 = \{1, ..., 40\}$, to sample back up to 40 observations. We maintained $N = 40$ particles and sampled $M = 1$ descendants of each type. The output of the particle filter can be seen in Figure 3. 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.

5 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 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 $W_t^{\frac{1}{2}} \sigma_I \nu_t = 10$.



(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

Figure 4: 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 H: [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.

2. The random walk model with two measurements

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

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, we set $\left(V_t^{(1)}\right)^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)} = 10$ or $\left(V_t^{(2)}\right)^{\frac{1}{2}} \sigma_A^{(2)} \epsilon_t^{(2)} = 10$ and for innovative outliers, we set $W_t^{\frac{1}{2}} \sigma_I \nu_t = 10$.

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 we set $V_t^{\frac{1}{2}} \epsilon_t = 30$ and the innovative outliers were instances where we set $\left(W_t^{(1)}\right)^{\frac{1}{2}} \eta_t^{(1)} = 100$ or $\left(W_t^{(2)}\right)^{\frac{1}{2}} \eta_t^{(2)} = 500$.

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

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

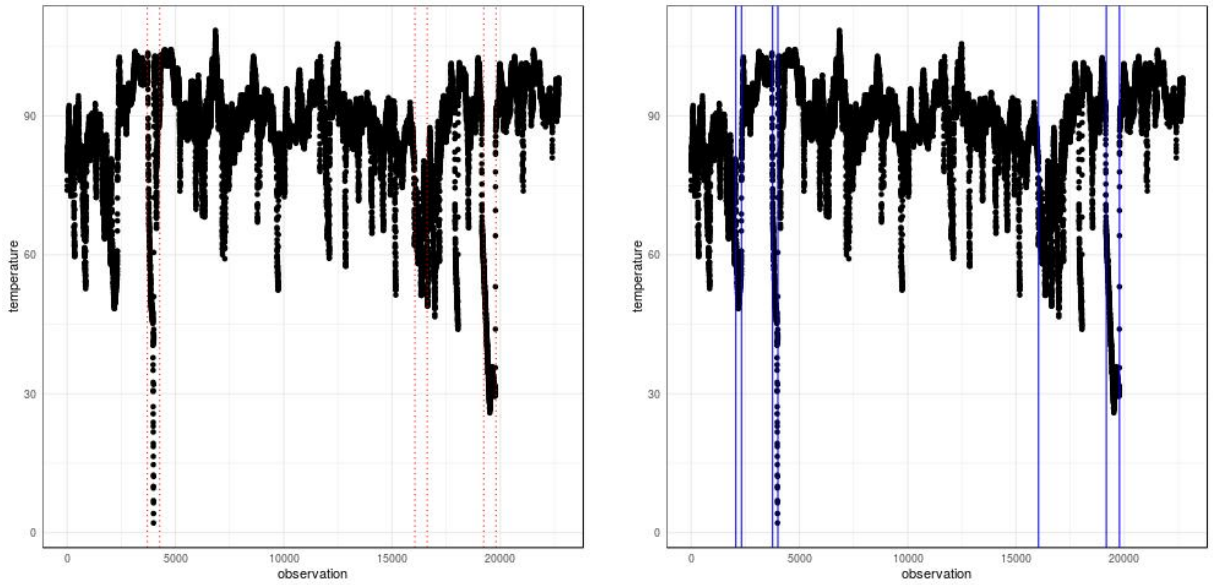
where $\sigma_A^{(1)} = \sigma_A^{(2)} = 1$. We consider a case with only additive outliers (in the first component only), a case with only innovative outliers (one in the second component, then one in the first, then one in the second, then one in the first), and a case with an additive outlier at time $t = 100$, 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, we set $\left(V_t^{(1)}\right)^{\frac{1}{2}} \sigma_A^{(1)} \epsilon_t^{(1)} = 30$ and for innovative outliers, we set $\left(W_t^{(1)}\right)^{\frac{1}{2}} \sigma_I^{(1)} \eta_t^{(1)} = 100$ or $\left(W_t^{(2)}\right)^{\frac{1}{2}} \sigma_I^{(2)} \eta_t^{(2)} = 500$.

We evaluate the different methods based on average predictive log-likelihood and average predictive mean squared error. 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, as the MSE is only an appropriate metric for unimodal posteriors.

The average log-likelihoods across all models can be found in Figure 4, while the qualitatively very similar results for the mean squared error can be found in the appendix. 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.

6 Application

In this section, we apply CE-BASS to two real datasets. We will use different types of models for the two applications to illustrate the way in which CE-BASS can be used. 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 emulates a changepoint model. The second is



(a) Raw data with labels

(b) CE-BASS output

Figure 5: Machine temperature dataset. The labelled anomalies are: a planned shutdown, an early warning sign of a problem, and the catastrophic system failure caused by the problem.

an unlabelled dataset which consist 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.

6.1 Machine Temperature Data

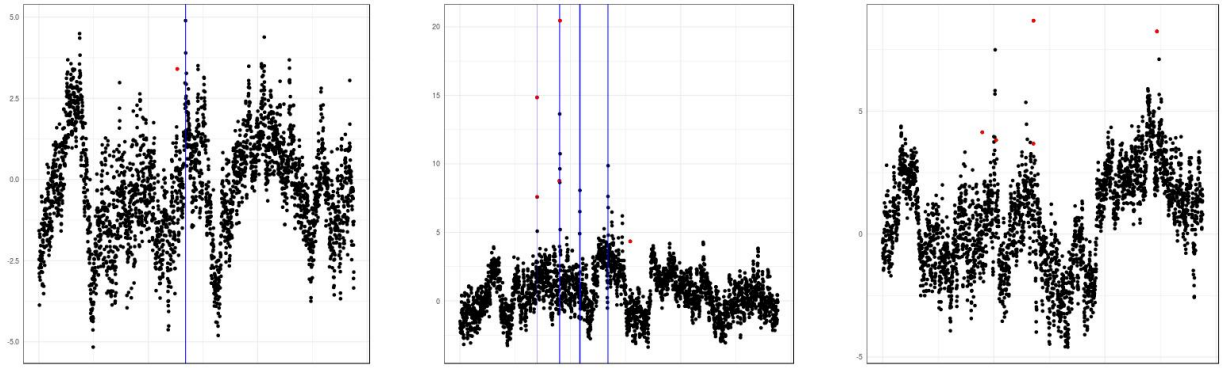
We now apply CE-BASS to the machine temperature data taken from the Numenta Anomaly Benchmark (NAB, [21]) which can be accessed at *https://github.com/numenta/NAB*. The data consists of over 20000 readings from a temperature sensor on a large industrial machine and is displayed in Figure 5a along the 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 by setting $\mathcal{B}_1 = \{1, 5, 10, 20, 40, 80, 150, 250\}$ 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 [21] as train 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 $\frac{1}{1-\rho}$.

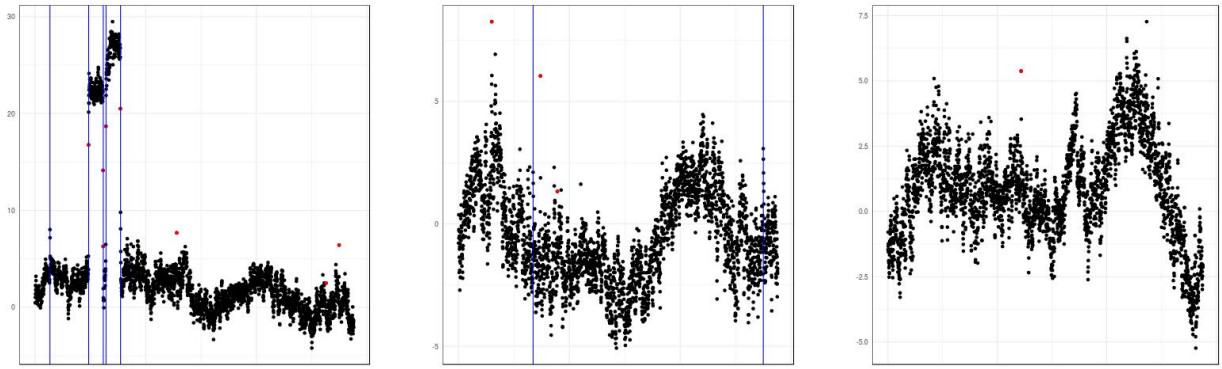
The results of this analysis can be seen in Figure 5b. We note that all anomalies flagged by the engineer are also being detected by CE-BASS. Two additional innovative anomalies around a prolonged drop which preceded the planned shutdown are also detected. They could be a false positive or an early warning sign of an anomaly prevented by the shutdown which has not been noticed by the engineer.

6.2 Router Data

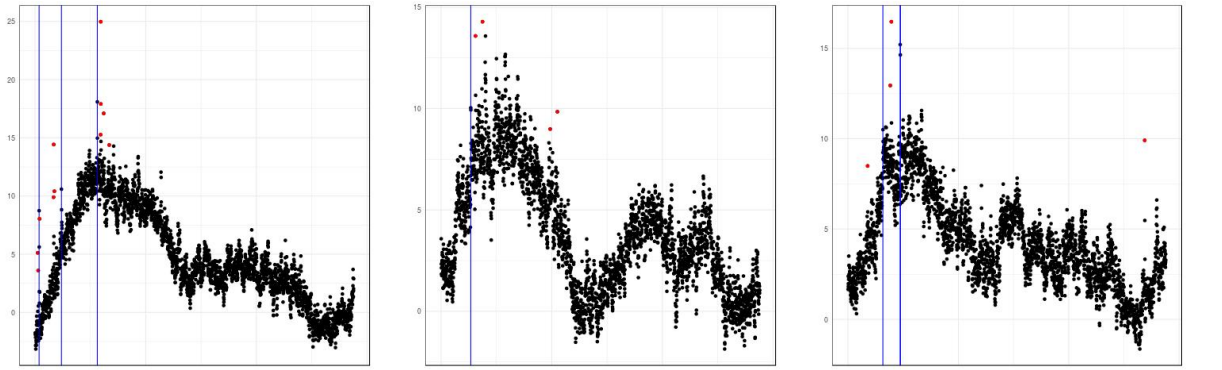
The online analysis of aggregated traffic data on servers is an important challenge in both predictive maintenance and cyber security. This is because anomalies in throughput can point towards problems in the network such as malfunctions or malicious behaviour. Detecting anomalies as soon as possible therefore means that the root cause can be addressed more quickly – potentially even before user experience is affected or harm caused.



(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

Figure 6: CE-BASS applied to 9 days of de-seasonalised router data. Lines correspond to innovative anomalies, i.e. spikes or level shifts.

In this section, we consider 19 days worth of data from a network IP router which has been 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 6a to 6i. 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

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

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 obtain the estimates $\sigma_A = 0.0516$, $\sigma_I^{(1)} = 0.0157$, $\sigma_I^{(2)} = 0.516$, and $\rho = 0.815$. The result obtained from running CE-BASS with these parameters on the daily router data is displayed in Figures 6a to 6i. We note that very few of the anomalies returned can be classed as false positives. At the same time, 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.

7 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.

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[2] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

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

[4] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[5] Mital A Gandhi and Lamine Mili. Robust Kalman filter based on a generalized maximum-likelihood-type estimator. *IEEE Transactions on Signal Processing*, 58(5):2509–2520, 2009.

[6] Yulong Huang, Yonggang Zhang, Ning Li, Zhemin Wu, and Jonathon A Chambers. A novel robust student's t-based Kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, 53(3):1545–1554, 2017.

[7] Jo-Anne Ting, Evangelos Theodorou, and Stefan Schaal. Learning an outlier-robust Kalman filter. In *European Conference on Machine Learning*, pages 748–756. Springer, 2007.

[8] Gabriel Agamennoni, Juan I Nieto, and Eduardo M Nebot. An outlier-robust Kalman filter. In *2011 IEEE International Conference on Robotics and Automation*, pages 1551–1558. IEEE, 2011.

[9] Peter Ruckdeschel, Bernhard Spangl, and Daria Pupashenko. Robust Kalman tracking and smoothing with propagating and non-propagating outliers. *Statistical Papers*, 55(1):93–123, 2014.

[10] Guobin Chang. Robust Kalman filtering based on Mahalanobis distance as outlier judging criterion. *Journal of Geodesy*, 88(4):391–401, 2014.

[11] Yulong Huang, Yonggang Zhang, Yuxin Zhao, and Jonathon A Chambers. A novel robust gaussian-student's t mixture distribution based Kalman filter. *IEEE Transactions on Signal Processing*, 2019.

[12] Genshiro Kitagawa. Non-gaussian state—space modeling of nonstationary time series. *Journal of the American statistical association*, 82(400):1032–1041, 1987.

[13] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2):107–113, 1993.

[14] Paul Fearnhead and Hans R. Künsch. Particle filters and data assimilation. *Annual Review of Statistics and Its Application*, 5(1):421–449, 2018.

[15] Rong Chen and Jun S Liu. Mixture Kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):493–508, 2000.

[16] Michael K Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.

[17] Paul Fearnhead and Guillem Rigaill. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183, 2019.

[18] Hyeyoung Maeng and Piotr Fryzlewicz. Detecting linear trend changes and point anomalies in data sequences. *arXiv preprint arXiv:1906.01939*, 2019.

[19] Paul Fearnhead and Peter Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.

[20] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on signal processing*, 50(2):174–188, 2002.

[21] Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 38–44. IEEE, 2015.

8 Appendix

8.1 Theorems and Derivations

8.1.1 Theorem 1

Theorem 1 *Let the prior for the hidden state $\boldsymbol{X}_t$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and an observation $\boldsymbol{Y}_{t+1} := \boldsymbol{Y}$ be available. Then the samples for $\tilde{\boldsymbol{V}}_{t+1}^{(i,i)}$ from*

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

have associated weight

$$\frac{1}{M} r_i \frac{\Gamma(a_i + \frac{1}{2})}{\Gamma(a_i)} \sqrt{\tilde{\sigma}_i} \frac{a_i^{a_i}}{\left(a_i + \frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\boldsymbol{Y}-\boldsymbol{CA\mu})}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right)^{a_i+\frac{1}{2}}} \frac{\exp\left(-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{CA\mu})^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\boldsymbol{CA\mu})\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\boldsymbol{V}}^{(i,i)} + \boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right)}}$$
$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\boldsymbol{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2 \frac{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} + \tilde{\boldsymbol{V}}_{t+1}^{(i,i)}}\right)\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}(\boldsymbol{Y}-\boldsymbol{CA\mu})}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\right)^2\right).$$

Proof: We wish to sample from the posterior distribution of $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ which is proportional to

$$r_i f_i\left(\tilde{\mathbf{V}}_{t+1}^{(i,i)}\right) \frac{\exp\left(-\frac{1}{2}(\mathbf{Y}-\mathbf{CA\mu})^T\left(\hat{\boldsymbol{\Sigma}} + \frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\mathbf{I}^{(i)}\right)^{-1}(\mathbf{Y}-\mathbf{CA\mu})\right)}{\sqrt{\left|\hat{\boldsymbol{\Sigma}} + \frac{\boldsymbol{\Sigma}_A^{(i,i)}}{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}\mathbf{I}^{(i)}\right|}}, \tag{6}$$

where $f_i()$ denotes the PDF of a $\tilde{\sigma}_i\Gamma(a_i, a_i)$-distribution. The intractable part in the above consists of

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

where $\mathbf{I}^{(i)} = \mathbf{e}_i\mathbf{e}_i^T$ is a matrix which is 0 everywhere with the exception of the ith entry of the ith row, which is 1. Note that $\mathbf{I}^{(i)}$ has rank 1 and therefore, by the Sherman Morrison formula,

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

Furthermore, given $tr(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(i)}) = \left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}$, the above is equal to

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

Crucially, the first term is constant in $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$, while the second is linear in $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ and therefore conjugate to the prior of $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$. The last term is quadratic in $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ and therefore vanishing much faster than the other two terms as $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ goes to 0, i.e. as the anomaly becomes stronger.

A very similar result for rank 1 updates of determinants, the matrix determinant Lemma, can be used to show that

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

Furthermore, given that

$$-\frac{1}{2}\left(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu}\right)^T \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{I}^{(j)}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu}\right)$$

is equal to

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

we can rewrite the posterior of $\tilde{\mathbf{V}}_{t+1}^{(i,i)}$ in Equation (6) as

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

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

Using conjugacy, we can therefore sample M particles for $\tilde{\mathbf{V}}^{(i,i)}$ from

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

and give each particle an importance weight proportional to

$$\frac{1}{M}r_i\frac{\Gamma(a_i + \frac{1}{2})}{\Gamma(a_i)}\sqrt{\tilde{\sigma}_i}\frac{a_i^{a_i}}{\left(a_i + \frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu}\right)}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right)^{a_i + \frac{1}{2}}}\frac{\exp\left(-\frac{1}{2}\left(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu}\right)^T\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu}\right)\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\mathbf{V}}^{(i,i)} + \boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\frac{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} + \tilde{\mathbf{V}}_{t+1}^{(i,i)}}\right)\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\left(\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu}\right)}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\right)^2\right).$$

8.1.2 Theorem 2

Theorem 2 *Let the prior for the hidden state $\boldsymbol{X}_t$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and an observation $\boldsymbol{Y}_{t+1} := \boldsymbol{Y}$ be available. Then the samples for $\tilde{\boldsymbol{W}}^{(j,j)}$ from*

$$\hat{\sigma}_i\Gamma\left(b_j + \frac{1}{2}, b_j + \frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\boldsymbol{C}^T\right)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{CA\mu}\right)}{\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}\right)^2\right)$$

have associated weight

$$\frac{1}{M}s_j\frac{\Gamma(b_i + \frac{1}{2})}{\Gamma(b_j)}\sqrt{\hat{\sigma}_j}\frac{b_j^{b_j}}{\left(b_j + \frac{\hat{\sigma}_i}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\boldsymbol{C}^T\right)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{CA\mu}\right)}{\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}\right)^2\right)^{b_i + \frac{1}{2}}}\frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{CA\mu}\right)^T\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{CA\mu}\right)\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\boldsymbol{W}}^{(j,j)} + \boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\boldsymbol{W}}^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}\right)^2\frac{\boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)} + \tilde{\boldsymbol{W}}_{t+1}^{(j,j)}}\right)\left(\frac{\left(\boldsymbol{C}^T\right)^{(j,:)}\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{CA\mu}\right)}{\sqrt{\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}\right)^{(j,j)}}}\right)^2\right)$$

The proof is almost identical to that of Theorem 1 and has been omitted.

8.1.3 Theorem 3

Theorem 3 *Let the prior for the hidden state $\boldsymbol{X}_t$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and an observation $\boldsymbol{Y}_{t+1} := \boldsymbol{Y}$ be available. Then the proposal particle $(\boldsymbol{I}_p, \boldsymbol{I}_q)$ for $(\boldsymbol{V}_t, \boldsymbol{W}_t)$ has weight proportional to*

$$(1 - \sum_{i=1}^{p} r_i - \sum_{j=1}^{q} s_j) \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{CA\mu}\right)^T \hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{CA\mu}\right)\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}}.$$

This is immediate from the Gaussian likelihood and the Bernoulli priors for $\lambda_t^{(i)}$ and $\gamma_t^{(j)}$.

8.1.4 Theorem 4

Theorem 4 *Let the prior for the hidden state $\boldsymbol{X}_t$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and an observation $\boldsymbol{Y}_{t+1} := \boldsymbol{Y}$ be available. When*

$$\tilde{\sigma}_i = \boldsymbol{\Sigma}_A^{(i,i)} \left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} \quad \textit{and} \quad \hat{\sigma}_j = \boldsymbol{\Sigma}_I^{(j,j)} \left(\boldsymbol{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{C}\right)^{(j,j)},$$

and $a_1 = ... = a_p = b_1 = ... = b_q = c$, the weights of additive and innovative anomalies are asymptotically proportional to

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

when

$$\boldsymbol{Y} - \boldsymbol{CA\mu} = \frac{\delta \boldsymbol{e}_i}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}} \quad \textit{and} \quad \boldsymbol{Y} - \boldsymbol{CA\mu} = \frac{\delta \boldsymbol{C}^{(:,j)}}{\sqrt{\left(\boldsymbol{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{C}\right)^{(j,j)}}},$$

respectively, as $\delta \to \infty$

Proof: Removing the likelihood term common to all particles the importance weights can be summarised as being

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

for the particles containing an anomaly in the ith additive component, and

$$\frac{1}{M} s_j \frac{\Gamma(b_i + \frac{1}{2})}{\Gamma(b_j)} \sqrt{\hat{\sigma}_j} \frac{b_j^{b_j}}{\left(b_j + \frac{\hat{\sigma}_i}{2\boldsymbol{\Sigma}_I^{(j,j)}} \left(\frac{\left(\mathbf{C}^T\right)^{(j,:)} \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{Y} - \mathbf{CA\mu})}{\left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)}}\right)^2\right)^{b_i + \frac{1}{2}}} \frac{1}{\sqrt{\left(\tilde{\mathbf{W}}^{(j,j)} + \boldsymbol{\Sigma}_I^{(j,j)} \left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)}\right)}}$$
$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\mathbf{W}}^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)} \left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)}}\right)^2 \frac{\boldsymbol{\Sigma}_I^{(j,j)} \left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)} \left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)} + \tilde{\mathbf{W}}_{t+1}^{(j,j)}}\right)\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)} \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{Y} - \mathbf{CA\mu})}{\sqrt{\left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{C}\right)^{(j,j)}}}\right)^2\right)$$

for the particles containing an anomaly in the jth innovative component.

As mentioned in Section II that the mean of the proposal of the ith additive component behaves asymptotically as

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

Furthermore, the standard deviation is on the same scale. We therefore have that

$$\tilde{\mathbf{V}}_{t+1}^{(i,i)} \sim \frac{1}{\delta^2}$$

as $\delta \to \infty$. The weight of an anomaly in the ith additive component therefore asymptotically behaves as

$$\frac{a_i^{a_i} \frac{1}{M} r_i \frac{\Gamma(a_i+\frac{1}{2})}{\Gamma(a_i)} \exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\delta^2\right)^{a_i}}$$

when $\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu} = \frac{1}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\delta\mathbf{e}_i$ as $\delta \to \infty$. A very similar reasoning can be used to show that the weight of an anomaly in the jth innovative component converges to

$$\frac{b_j^{b_j} \frac{1}{M} s_j \frac{\Gamma(b_j+\frac{1}{2})}{\Gamma(b_j)} \exp\left(\frac{1}{2}\delta^2\right)}{\left(\frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\delta^2\right)^{b_j}}$$

when $\mathbf{Y} - \mathbf{CA}\boldsymbol{\mu} = \frac{\mathbf{C}^{(:,ij}}{\sqrt{\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}}\delta$ as $\delta \to \infty$.

The result then follows when all the b_js and the a_is are equal to the same constant c and

$$\tilde{\sigma}_i = \boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} \quad \text{and} \quad \hat{\sigma}_j = \boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}.$$

8.1.5 Theorem 5

Theorem 5 *Let the prior for the hidden state $\boldsymbol{X}_{t-k}$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the samples for $\tilde{\boldsymbol{W}}_{t-k+1}^{(j,j)}$ from*

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

where $\tilde{\boldsymbol{z}}_{t+1-k}^{(k)} = \tilde{\boldsymbol{Y}}_{t+1-k}^{(k)} - \tilde{\boldsymbol{C}}^{(k)}\boldsymbol{A}\boldsymbol{\mu}$ have associated weight

$$\frac{\frac{1}{M}s_i\left(1-\sum_{i'=1}^p r_{i'}-\sum_{j'=1}^q s_{j'}\right)^k \frac{\Gamma(b_j+\frac{1}{2})}{\Gamma(b_j)}\sqrt{\hat{\sigma}_j}b_j^{b_j}}{\left(b_i + \frac{\hat{\sigma}_j}{2\boldsymbol{\Sigma}_I^{(j,j)}}\left(\frac{\left(\left(\tilde{\boldsymbol{c}}^{(k)}\right)^T\right)^{(j,:)}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{z}}_{t+1-k}^{(k)}\right)}{\left(\left(\tilde{\boldsymbol{c}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\tilde{\boldsymbol{c}}^{(k)}\right)^{(j,j)}}\right)^2\right)^{b_j+\frac{1}{2}}} \frac{\exp\left(-\frac{1}{2}\left(\tilde{\boldsymbol{z}}_{t+1-k}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{z}}_{t+1-k}^{(k)}\right)\right)}{\sqrt{\left|\hat{\boldsymbol{\Sigma}}^{(k)}\right|}\sqrt{\left(\boldsymbol{W}^{(j,j)} + \boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1+\left(\frac{\boldsymbol{W}_{t+1}^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}}\right)^2 \frac{\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}}{\boldsymbol{\Sigma}_I^{(j,j)}\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)} + \boldsymbol{W}_{t+1}^{(j,j)}}\right)$$

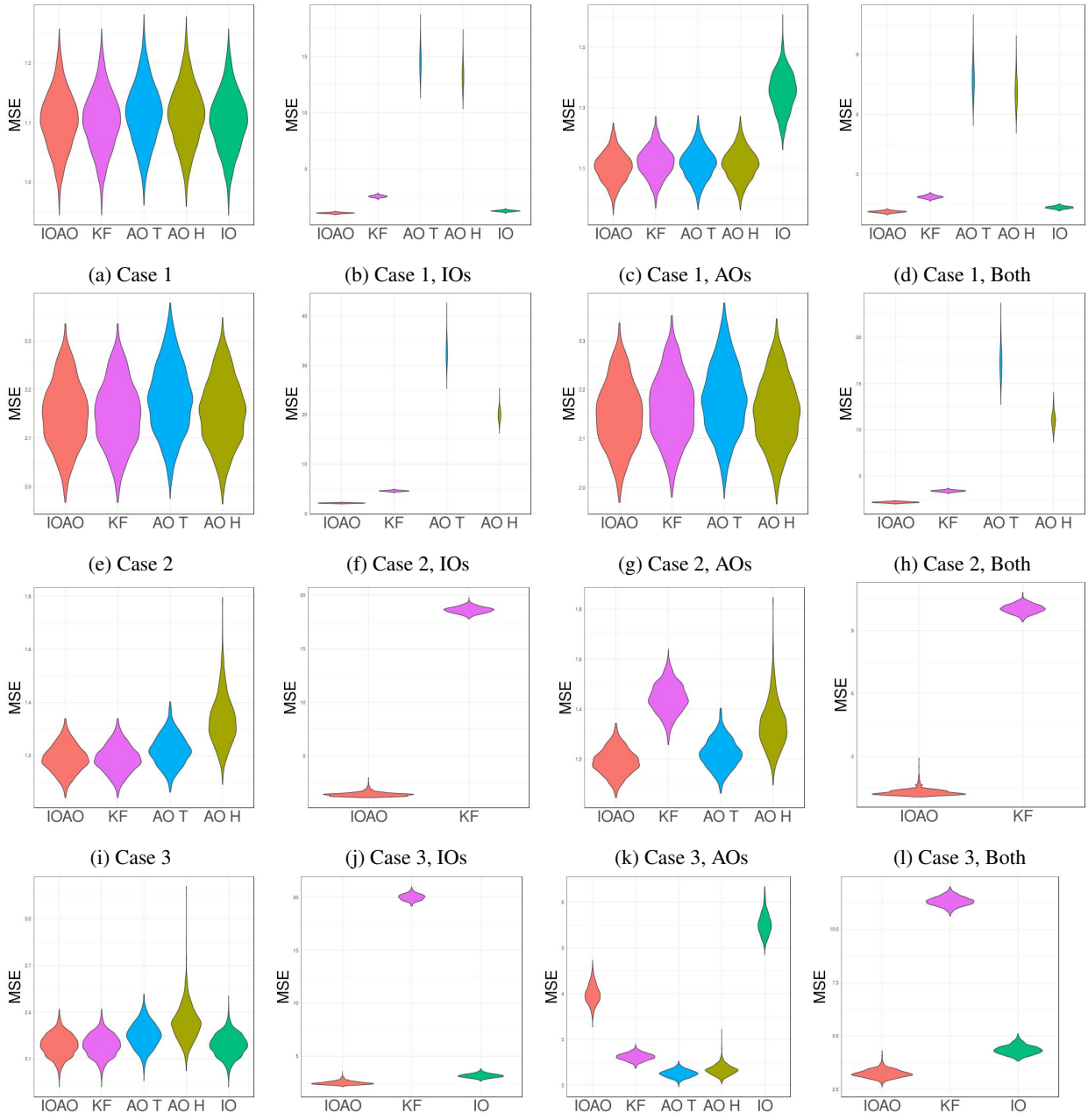
$$\left(\frac{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\right)^{(j,:)}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{Y}}_{t+1-k}^{(k)} - \left(\tilde{\boldsymbol{C}}^{(k)}\right)\boldsymbol{A}\boldsymbol{\mu}_{t-k}\right)}{\sqrt{\left(\left(\tilde{\boldsymbol{C}}^{(k)}\right)^T\left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\left(\tilde{\boldsymbol{C}}^{(k)}\right)\right)^{(j,j)}}}\right)^2\Bigg)$$

Proof: Identical (up to variable names) to that of Theorem 2.

8.2 Additional Simulations

Violin plots for the predictive mean squared error are displayed in Figure 7

8.3 Complete pseudocode



(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

Figure 7: Violin plots for the average predictive mean squared error of the five filters over the four different scenarios under a range of models. Lower values correspond to better performance. Methods are omitted if they can not be applied to the setting or if their performance is too poor.

Algorithm 3 KF_Upd($\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I$)

$\boldsymbol{\mu}_p \leftarrow \mathbf{A}\boldsymbol{\mu}$
$\boldsymbol{\Sigma}_p \leftarrow \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Sigma}_I$
$\mathbf{z} = \mathbf{Y} - \boldsymbol{\mu}_p$
$\hat{\boldsymbol{\Sigma}} \leftarrow \mathbf{C}\boldsymbol{\Sigma}_p\mathbf{C}^T + \boldsymbol{\Sigma}_A$
$\mathbf{K} \leftarrow \boldsymbol{\Sigma}_p\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}$
$\boldsymbol{\mu}_{new} \leftarrow \boldsymbol{\mu}_p + \mathbf{K}\mathbf{z}$
$\boldsymbol{\Sigma}_{new} \leftarrow (\mathbf{I} - \mathbf{K}\mathbf{C})\,\boldsymbol{\Sigma}_p$

Output: $(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$

Algorithm 4 Sample_typical($\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I$)

$\mathbf{V} \leftarrow \mathbf{I}_p$
$\mathbf{W} \leftarrow \mathbf{I}_q$
$\hat{\boldsymbol{\Sigma}} \leftarrow \mathbf{C}\left(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Sigma}_I\right)\mathbf{C}^T + \boldsymbol{\Sigma}_A$
$\mathbf{z} \leftarrow \mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}$
$prob \leftarrow \left(1 - \sum_{i=1}^{p} r_i - \sum_{j=1}^{q} s_j\right)\exp\left(-\frac{1}{2}\mathbf{z}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{z}\right)/\sqrt{\left|\hat{\boldsymbol{\Sigma}}\right|}$

Output: $(\mathbf{V}, \mathbf{W}, prob)$

Algorithm 5 Sample_add_comp($i, \mathbf{z}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}_A, M$)

$\mathbf{V} \leftarrow \mathbf{I}_p$
$\mathbf{V} \leftarrow \mathbf{I}_q$
$\mathbf{V}^{(i,i)} \leftarrow \tilde{\sigma}_i\Gamma\left(a_i + \frac{1}{2}, a_i + \frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\mathbf{z}}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right)$

$$prob \leftarrow \frac{1}{M}r_i\frac{\Gamma(a_i+\frac{1}{2})}{\Gamma(a_i)}\frac{a_i^{a_i}}{\left(a_i + \frac{\tilde{\sigma}_i}{2\boldsymbol{\Sigma}_A^{(i,i)}}\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\mathbf{z}}{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\right)^{a_i+\frac{1}{2}}}\frac{\sqrt{\tilde{\sigma}_i}\exp\left(-\frac{1}{2}\mathbf{z}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{z}\right)}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}\sqrt{\left(\tilde{\mathbf{V}}^{(i,i)} + \boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}\right)}}$$
$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\mathbf{V}}_{t+1}^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}\right)^2\frac{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}{\boldsymbol{\Sigma}_A^{(i,i)}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)} + \tilde{\mathbf{V}}_{t+1}^{(i,i)}}\right)\left(\frac{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,:)}\mathbf{z}}{\sqrt{\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)^{(i,i)}}}\right)^2\right).$$

Output: $(\mathbf{V}, \mathbf{W}, prob)$

Algorithm 6 Sample_add($\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M$)

$\hat{\boldsymbol{\Sigma}} \leftarrow \mathbf{C}\left(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Sigma}_I\right)\mathbf{C}^T + \boldsymbol{\Sigma}_A$
$\mathbf{z} \leftarrow \mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}$
$Add\_Pt \leftarrow \{\}$ ▷ Additive Anom. Particles
for $i \in \{1, ..., p\}$ **do**
    $Add\_Pt \leftarrow Add\_Pt \cup \{\text{Sample\_add\_comp}(i, \mathbf{z}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}_A, M)\}$
end for

Output: Add_Pt

Algorithm 7 Sample_inn_comp($j, \mathbf{z}, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_I, M$)

$\mathbf{V} \leftarrow \mathbf{I}_p$

$\mathbf{V} \leftarrow \mathbf{I}_q$

$\mathbf{W}^{(i,i)} \leftarrow \hat{\sigma_i}\Gamma\left(b_i + \frac{1}{2}, b_i + \frac{\hat{\sigma_i}}{2\mathbf{\Sigma}_I^{(i,i)}}\left(\frac{\left(\mathbf{C}^T\right)^{(i,:)}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}}{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(i,i)}}\right)^2\right)$

$$prob \leftarrow \frac{1}{M}s_j\frac{\Gamma(b_i+\frac{1}{2})}{\Gamma(b_j)}\frac{b_j^{b_j}}{\left(b_j + \frac{\hat{\sigma_i}}{2\mathbf{\Sigma}_I^{(j,j)}}\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}}{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2\right)^{b_i+\frac{1}{2}}}\frac{\sqrt{\hat{\sigma}_j}\exp\left(-\frac{1}{2}\mathbf{z}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}\right)}{\sqrt{|\hat{\mathbf{\Sigma}}|}\sqrt{\left(\tilde{\mathbf{W}}^{(j,j)} + \mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}\right)}}$$

$$\exp\left(\frac{1}{2}\left(1 + \left(\frac{\tilde{\mathbf{W}}^{(j,j)}}{\mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}\right)^2\frac{\mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}{\mathbf{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)} + \tilde{\mathbf{W}}_{t+1}^{(j,j)}}\right)\left(\frac{\left(\mathbf{C}^T\right)^{(j,:)}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}}{\sqrt{\left(\mathbf{C}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}}}\right)^2\right)$$

Output: $(\mathbf{V}, \mathbf{W}, prob)$

Algorithm 8 Sample_inn($\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I, M$)

$\hat{\mathbf{\Sigma}} \leftarrow \mathbf{C}\left(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{\Sigma}_I\right)\mathbf{C}^T + \mathbf{\Sigma}_A$

$\mathbf{z} \leftarrow \mathbf{Y} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}$

$Inn\_Pt \leftarrow \{\}$ ▷ Innovative Anom. Particles

for $i \in \{1, ..., q\}$ **do**

  $Inn\_Pt \leftarrow Inn\_Pt \cup \{\text{Sample\_inn\_comp}(i, \mathbf{z}, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_I, M)\}$

end for

Output: Inn_Pt

Algorithm 9 Sample_Particles($M, \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I$)

$Desc \leftarrow \{\}$ ▷ To store Descendants

$Desc \leftarrow Desc \cup \text{Sample\_typical}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I)$

for $i \in 1, ..., M$ **do**

  $Desc \leftarrow Desc \cup \text{Sample\_add}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I, M)$

end for

for $i \in 1, ..., M$ **do**

  $Desc \leftarrow Desc \cup \text{Sample\_inn}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{Y}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I, M)$

end for

Output: $Desc$

Algorithm 10 BS_inn ($\boldsymbol{\mu}, \mathbf{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \mathbf{\Sigma}_A, \mathbf{\Sigma}_I, M, horizon$)

$\tilde{\mathbf{C}} \leftarrow \mathbf{C}\left[\left(\mathbf{A}^0\right)^T, ..., \left(\mathbf{A}^{horizon}\right)^T\right]^T$

$\tilde{\mathbf{z}} \leftarrow \tilde{\mathbf{Y}} - \tilde{\mathbf{C}}\mathbf{A}\boldsymbol{\mu}$

$\tilde{\mathbf{\Sigma}} \leftarrow \tilde{\mathbf{C}}\left(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{I}_{horizon} \otimes \mathbf{\Sigma}_I\right)\tilde{\mathbf{C}}^T + \mathbf{I}_{horizon} \otimes \mathbf{\Sigma}_A$

$Cd \leftarrow \{\}$ ▷ To store Candidates.

for $i \in \{1, .., q\}$ **do**

  if $horizon \in \mathcal{B}_i$ **then**

    for $j \in \{1, ..., M\}$ **do**

      $Cd \leftarrow Cd \cup \{\text{Sample\_inn\_comp}(i, \tilde{\mathbf{z}}, \tilde{\mathbf{\Sigma}}, \mathbf{A}, \tilde{\mathbf{C}}, \mathbf{\Sigma}_I, M \cdot |\mathcal{B}_i|)\}$

    end for

  end if

end for

Output: $Cand$

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 $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\}$

for $t \in \mathbb{N}^+$ **do**
  $Candidates \leftarrow \{\}$
  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in Particles(t-1)$ **do**
    $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_Particles$(M, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
    $Candidates \leftarrow Candidates \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob)\}$
  end for
  $Descendants \leftarrow$ Subsample$(N, Candidates)$
  $Particles(t) \leftarrow \{\}$
  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob) \in Descendants$ **do**
    $(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new}) \leftarrow$ KF_Upd$(\mathbf{Y}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$
    $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})\}$
  end for
end for

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 $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, 1)\}$
Set $max\_horizon = \max(\cup_{i=1}^{q} \mathcal{B}_i)$

for $t \in \mathbb{N}^+$ **do**
  $Cand \leftarrow \{\}$ ▷ To Store Candidates
  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-1)$ **do**
    $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_typical$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
    $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$
    $Add\_Des \leftarrow$ Sample_add$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_t, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$
    for $(\mathbf{V}, \mathbf{W}, prob) \in Add\_Des$ **do**
      $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, 1)\}$
    end for
  end for
  for $hor \in \{1, ..., max\_horizon\}$ **do**
    for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob_{prev}) \in Particles(t-hor)$ **do**
      $\tilde{\mathbf{Y}} \leftarrow \left[\mathbf{Y}_{t-hor+1}^T, ..., \mathbf{Y}_t^T\right]^T$
      $Inn\_Des \leftarrow$ BS_inn$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M, hor)$
      for $(\mathbf{V}, \mathbf{W}, prob) \in Inn\_Des$ **do**
        $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot prob_{prev}, hor)\}$
      end for
    end for
  end for
  $Desc \leftarrow$ Subsample$(N, Cand)$ ▷ Sampling proportional to $prob$
  $Particles(t) \leftarrow \{\}$
  for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, hor) \in Desc$ **do**
    $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+1-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}^{1/2}\boldsymbol{\Sigma}_A, \mathbf{W}^{1/2}\boldsymbol{\Sigma}_I)$
    if $hor > 1$ **then**
      for $i \in \{2, ..., hor\}$ **do**
        $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+i-hor}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
      end for
    end if
    $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, prob \cdot \frac{|Cand|}{|Desc|})\}$
  end for
end for