# Innovative And Additive Outlier Robust Kalman Filtering With A Robust Particle Filter – Supplementary Material

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

## I. Choice of Parameters with Back-Sampling

Here we describe how to choose hyper-parameters to make the algorithm robust. As in Section III-B, 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 $i$th innovative component at time $t-k+1$, i.e.

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

As $\delta$ grows, the importance weight behaves as

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

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

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

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

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

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

## II. Theorems and Derivations

### A. Theorem 1

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

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

have associated weight

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

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

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

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

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

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

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

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

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

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

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

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

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

Furthermore, given that

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

is equal to

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

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

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

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

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

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

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

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

and give each particle an importance weight proportional to

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

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

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

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

### *B. Theorem 2*

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

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

have associated weight

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

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

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

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

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

### *C. Theorem 3*

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

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

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

### D. Theorem 4

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

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

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

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

when

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

respectively, as $\delta \to \infty$

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

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

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

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

for the particles containing an anomaly in the $i$th additive component, and

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

$$\frac{1}{\sqrt{\left(\tilde{\mathbf{W}}^{(j,j)} + \boldsymbol{\Sigma}_I^{(j,j)}\left(\mathbf{C}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{C}\right)^{(j,j)}\right)}}$$

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

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

for the particles containing an anomaly in the $j$th innovative component.

As mentioned in Section II that the mean of the proposal of the $i$th additive component behaves asymptotically as

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

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 $i$th 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 $j$th 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 $b_j$s and the $a_i$s are equal to the same constant $c$ and

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

### E. Theorem 5

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

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

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

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

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

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

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

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

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

## III. Transforming Model to have Diagonal Covariance Matrix

If we have a general state space model

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

where $\mathbf{\Sigma}_A^{\frac{1}{2}}$ and/or $\mathbf{\Sigma}_I^{\frac{1}{2}}$ are not diagonal, we can apply a linear transformation so as to obtain an equivalent model where these covariance matrices are diagonal.

For the following assume that $\mathbf{\Sigma}_A$ and $\mathbf{\Sigma}_I$ are of full-rank. If they are not then we can obtain an equivalent model after removing one or more components of the noise $\boldsymbol{\epsilon}_t$ or $\boldsymbol{\nu}_t$. As $\mathbf{\Sigma}_A$ and $\mathbf{\Sigma}_I$ are covariance matrices, they are symmetric and thus there exist matrices of orthogonal eigenvectors $\boldsymbol{P}_A$ and $\boldsymbol{P}_I$ and diagonal matrices of the square-root of the eigenvalues, $\boldsymbol{D}_A$ and $\boldsymbol{D}_I$, such that

$$\mathbf{\Sigma}_I = \boldsymbol{P}_I\boldsymbol{D}_I\boldsymbol{D}_I\boldsymbol{P}_I^T$$
$$\mathbf{\Sigma}_A = \boldsymbol{P}_A\boldsymbol{D}_A\boldsymbol{D}_A\boldsymbol{P}_A^T.$$

Thus we can choose $\mathbf{\Sigma}_I^{\frac{1}{2}} = \boldsymbol{P}_I\boldsymbol{D}_I$ and $\mathbf{\Sigma}_A^{\frac{1}{2}} = \boldsymbol{P}_A\boldsymbol{D}_A$.

Using the fact that $\boldsymbol{P}_I^T$ is the inverse of $\boldsymbol{P}_I$ as it is an orthogonal matrix, and similarly $\boldsymbol{P}_A^T$ is the inverse of $\boldsymbol{P}_A$, we can define a new state $\tilde{\mathbf{X}}_t = \boldsymbol{P}_I^T\mathbf{X}_t$ and a new observation $\tilde{\mathbf{Y}}_t = \boldsymbol{P}_A^T\mathbf{Y}_t$ which satisfy the model

$$\tilde{\mathbf{Y}}_t = \boldsymbol{P}_A^T\mathbf{C}\boldsymbol{P}_I\tilde{\mathbf{X}}_t + \boldsymbol{P}_A^T\mathbf{\Sigma}_A^{\frac{1}{2}}\mathbf{V}_t^{\frac{1}{2}}\boldsymbol{\epsilon}_t,$$
$$\tilde{\mathbf{X}}_t = \boldsymbol{P}_I^T\mathbf{A}\boldsymbol{P}_I\tilde{\mathbf{X}}_{t-1} + \boldsymbol{P}_I^T\mathbf{\Sigma}_I^{\frac{1}{2}}\mathbf{W}_t^{\frac{1}{2}}\boldsymbol{\nu}_t.$$

This is a state space model in the required form, with $\boldsymbol{P}_A^T\mathbf{\Sigma}_A^{\frac{1}{2}} = \boldsymbol{P}_A^T\boldsymbol{P}_A\boldsymbol{D}_A = \boldsymbol{D}_A$ diagonal. Similarly, $\boldsymbol{P}_I^T\mathbf{\Sigma}_I^{\frac{1}{2}}$ is diagonal.

## IV. Backsampling weights

Here we give an informal explanation of the form of weights for backsampling, and why they depend on the evidence, $p(\mathbf{Y}_{1:s})$ for $s = 1, \ldots,$.

To explain how we derive the backsampling weights let $\boldsymbol{\theta}_{1:t}$ denote a hidden state of our system up to time $t$, which we define to be the values of $\mathbf{V}_i$ and $\mathbf{W}_i$ for $i = 1, \ldots, t$. We can view each particle at time $t$ as corresponding to a specific realisation of $\boldsymbol{\theta}_{1:t}$ – though in practice we do not store all the historical values of $\mathbf{V}_i$ and $\mathbf{W}_i$ prior to time $t$, but instead store the value $\boldsymbol{\mu}_t$ and $\mathbf{\Sigma}_t$ which are the conditional mean and covariance of the state $\boldsymbol{x}_t$ given $\boldsymbol{\theta}_{1:t}$ and $\mathbf{Y}_{1:t}$.

For the backsampling algorithm it is helpful to think of proposing particles for $\boldsymbol{\theta}_{1:t+1}$. If we propose them with some proposal $q(\boldsymbol{\theta}_{1:t+1})$ then an appropriate IS weight will be proportional to

$$\frac{p(\boldsymbol{\theta}_{1:t+1}, \mathbf{Y}_{1:t+1})}{q(\boldsymbol{\theta}_{1:t+1})}.$$

The weighted samples from such an importance sampler will approximate a density proportional to $p(\boldsymbol{\theta}_{1:t+1}, \mathbf{Y}_{1:t+1}) \propto p(\boldsymbol{\theta}_{1:t+1}|\mathbf{Y}_{1:t+1})$.

For appropriate probabilities $p_1$, $p_2$ and $p_3$ with $p_1 + p_2 + p_3 = 1$ our proposal for backsampling is a mixture of the form:

(i) With probability $p_1$ we propose $\boldsymbol{\theta}_{1:t}$ from $p(\boldsymbol{\theta}_{1:t}|\mathbf{Y}_{1:t})$ and set $\boldsymbol{\theta}_{t+1}$ to correspond to no outlier.

(ii) With probability $p_2$ we propose $\boldsymbol{\theta}_{1:t}$ from $p(\boldsymbol{\theta}_{1:t}|\mathbf{Y}_{1:t})$ and set $\boldsymbol{\theta}_{t+1}$ to correspond to an additive outlier. The component is sampled uniformly, and then the value of $\mathbf{V}_{t+1}$ is sampled according to our proposal distribution.

(iii) With probability $p_3$ we sample $k$ uniformly from $1, \ldots, k_{\max}$. We then sample $\boldsymbol{\theta}_{1:t-k+1}$ from $p(\boldsymbol{\theta}_{1:t-k+1}|\mathbf{Y}_{1:t-k+1})$ and set $\boldsymbol{\theta}_{t-k+2}$ to be an innovative outlier, and, if $k > 1$, $\boldsymbol{\theta}_{t-k+3:t+1}$ to correspond to no outliers. The specific proposal for $\boldsymbol{\theta}_{t-k+2}$ is to choose a component uniformly at random and then the value of $\mathbf{W}_{t-k+2}$ is drawn from our proposal distribution.

As is standard in particle filtering, rather than sampling from the filtering densities, such as $p(\boldsymbol{\theta}_{1:t}|\mathbf{Y}_{1:t})$ in steps (i) and (ii), we sample from the particle approximation to these densities. In practice for appropriately chosen $p_1$, $p_2$ and $p_3$ we use a stratified sampling approach across components (i)–(iii) and across the type (component and, for innovative outliers, value of $k$) of outlier in steps (ii) and (iii). So $p_1$, $p_2$ and $p_3$ are each proportional to the number of proposals from (i), (ii) and (iii) respectively. (In the above $k_{\max}$ represents the maximum horizon in Algorithm 2.)

For (i) the proposal distribution is $p_1 p(\boldsymbol{\theta}_{1:t}|\mathbf{Y}_{1:t})$ times a point mass on $\boldsymbol{\theta}_{t+1}$ not being an outlier. Thus the importance sampling weight is

$$\frac{p(\boldsymbol{\theta}_{1:t}|\mathbf{Y}_{1:t})p(\mathbf{Y}_{1:t})p(\boldsymbol{\theta}_{t+1})p(\mathbf{Y}_{t+1}|\boldsymbol{\theta}_{1:t+1}, \mathbf{Y}_{1:t}}{p_1 p(\boldsymbol{\theta}_{1:t}|\mathbf{Y}_{1:t})}$$
$$= \quad p(\mathbf{Y}_{1:t})\frac{p(\boldsymbol{\theta}_{t+1})p(\mathbf{Y}_{t+1}|\boldsymbol{\theta}_{1:t+1}, \mathbf{Y}_{1:t})}{p_1 p(\boldsymbol{\theta}_{1:t}|\mathbf{Y}_{1:t})}.$$

It is equal to the evidence at time $t$, $p(\mathbf{Y}_{1:t})$ times the importance sampling weight for a non-outlier we have in the no-backsampling algorithm. For step(ii) we similarly get that the importance sampling weight is $p(\mathbf{Y}_{1:t})$ times the importance sampling weight for an additive outlier we have in the no-backsampling algorithm.

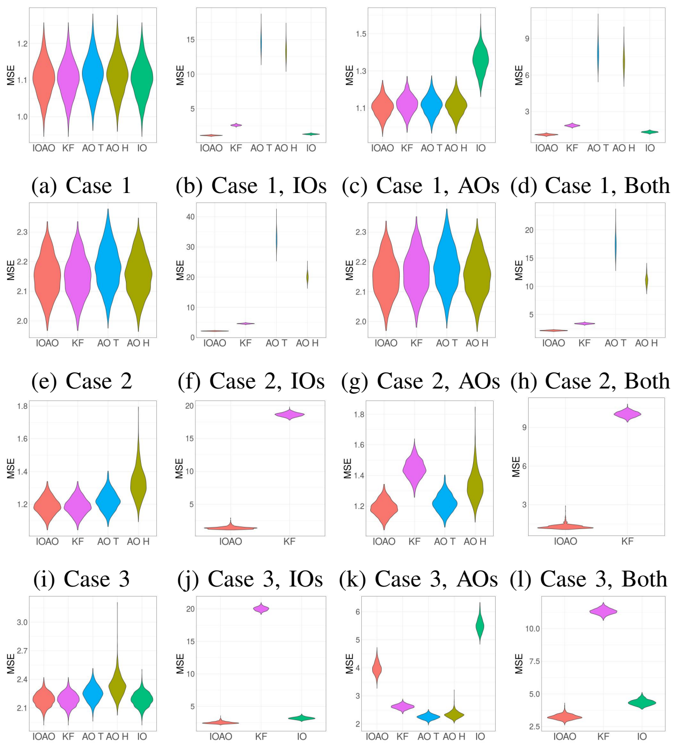
For step (iii) we have introduced an additional variable $k$, so we need to adapt the target to include this. This is a standard augmentation trick in importance sampling and we can do this by having a target where $k$ is independent and is drawn from a uniform distribution on $\{1, \ldots, k_{\max}\}$. If $q(\boldsymbol{\theta}_{t-k+2})$ is our proposal for $\boldsymbol{\theta}_{t-k+2}$ then the importance sampling weight is

$$\frac{p(k, \boldsymbol{\theta}_{1:t+1}, \mathbf{Y}_{1:t+1})}{q(\boldsymbol{\theta}_{1:t+1})}$$
$$= \frac{\frac{1}{k_{\max}}p(\boldsymbol{\theta}_{1:t-k+1}, \mathbf{Y}_{1:t-k+1})p(\boldsymbol{\theta}_{t-k+2:t+1}, \mathbf{Y}_{t-k+2:t+1}|\boldsymbol{\theta}_{1:t-k+1})}{p_3(1/k_{\max})p(\boldsymbol{\theta}_{1:t-k+1}|\mathbf{Y}_{1:t-k+1})q(\boldsymbol{\theta}_{t-k+2})}$$
$$= \quad p(\mathbf{Y}_{1:t-k+1})\frac{p(\boldsymbol{\theta}_{t-k+2:t+1}, \mathbf{Y}_{t-k+2:t+1}|\boldsymbol{\theta}_{1:t-k+1})}{p_3 q(\boldsymbol{\theta}_{t-k+2})}.$$

The term in the numerator can be obtained as

$$p(\boldsymbol{\theta}_{t-k+2:t+1}, \mathbf{Y}_{t-k+2:t+1}|\boldsymbol{\theta}_{1:t-k+1})$$
$$= \quad p(\boldsymbol{\theta}_{t-k+2:t+1})p(\mathbf{Y}_{t-k+2:t+1}|\boldsymbol{\theta}_{1:t+1}),$$

as the states (as we have defined) are independent at each time-point. The likelihood term $p(\mathbf{Y}_{t-k+2:t+1}|\boldsymbol{\theta}_{1:t+1})$ can be



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

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

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

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

Fig. 1: Violin plots for the average predictive mean squared error of the five filters (IOAO: CE-BASS, KF: The classical Kalman Filter, AO T: [1], AO H: [2], IO: [2]) 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.

calculated from $\boldsymbol{\theta}_{t-k+2:t+1}$ and the mean and covariance for the state (from the Kalman Filter) at time $t-k+1$.

The key thing to note is that for backsampling at horizon $k$ we have a factor $p(\mathbf{Y}_{1:t+1-k})$. In the standard particle filter algorithm $k=1$, and thus this term is the same as that which appears in the weights for (i) and (ii) – and thus can be ignored. With backsampling we need to include this term as we consider $k \neq 1$ as well – this explains the need to calculate the estimate of the evidence in Algorithm 2.

## V. Additional Simulations

Violin plots for the predictive mean squared error are displayed in Figure 1

In most cases these plots show similar performance to the plots of predictive log-likelihood shown in the paper. One exception is for Case 4 with additive outliers – where under mean square error CE-BASS performs much worse that the Kalman Filter and its versions which allow only for additive outliers. This is due to the fact that an additive oulier can be fit as either an additive outlier or an innovative outlier under the model used be CE-BASS. The predictive distribution for the observation at the next time-step is this bi-modal. Whilst the observation lies within one of the modes, and thus the filter is judged to perform well under the predictive log-likelihood

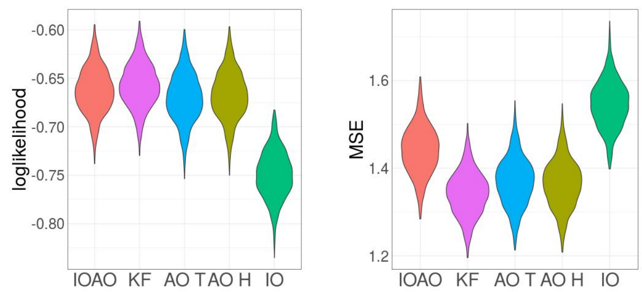



Fig. 2: Violin plots for the predictive log likelihood and the average predictive mean squared error of the five filters (IOAO: CE-BASS, KF: The classical Kalman Filter, AO T: [1], AO H: [2], IO: [2])

criteria, the mean of the predictive distribution can be far from the observation, and thus the mean square error is large.

We additionally analysed a scenario in which both additive and innovative anomalies occur at the same time. For this, we used the model of Example 1 with $\sigma_A = 1$ and $\sigma_A = 0.1$. We consider a case with an innovative and additive anomaly occur at times t = 300 and t = 600. To simulate additive anomalies, we set $V_t^{(1,1)} \sigma_A \epsilon_t = 10$ and to simulate the innovative outliers we set $W_t^{(1,1)} \sigma_I \eta_t = 10$. The resulting log-likelihood and MSE plots can be found in Figure 2. It is apparent from the log-likelihood plot that CE-BASS performs slightly worse than additive outlier filters, and even the Kalman filter. This is due to the fact it is unable to correctly capture the multi-modality around the anomaly – it can allow for there being either an additive or a innovative outlier, but not capture the mode that corresponds to both outliers occurring simultaneously.

## VI. Complete Pseudocode

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

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

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

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

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

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

## References


[1] G. Agamennoni, J. I. Nieto, and E. M. Nebot, "An outlier-robust Kalman filter," in *2011 IEEE International*

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

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

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

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

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

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

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

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

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

$Add\_Pt \leftarrow \{\}$ ▷ Additive Anom. Particles

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

    $Add_Pt \leftarrow Add_Pt \cup \{$Sample_add_comp$(i, \mathbf{z}, \hat{\mathbf{\Sigma}}, \mathbf{\Sigma}_A, M)\}$

**end for**

**Output:** $Add\_Pt$

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

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

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

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

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

**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**

    $Inn_Pt \leftarrow Inn_Pt \cup \{$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

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

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

    $Desc \leftarrow Desc \cup$ 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**

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

**end for**

**Output:** $Desc$

*Conference on Robotics and Automation*. IEEE, 2011, pp. 1551–1558.

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

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

**for** $i \in \{0, .., k\}$ **do**

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

**end for**

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

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

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

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

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

        $Cd \leftarrow Cd \cup \{$Sample_inn_comp$(i, \tilde{\mathbf{z}}, \tilde{\mathbf{\Sigma}}, \mathbf{A}, \tilde{\mathbf{C}}, \mathbf{\Sigma}_I, M)\}$

    **end for**

**end for**

**Output:** $Cd$

**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$(\mathbf{Y}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}\boldsymbol{\Sigma}_A, \mathbf{W}\boldsymbol{\Sigma}_I)$
    $Particles(t) \leftarrow Particles(t) \cup \{(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})\}$
  **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 $Particles(0) = \{(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, 1)\}$
$EV(t)$=1
Set $max\_horizon$

**for** $t \in \mathbb{N}$ **do**
  $Cand \leftarrow \{\}$
  **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in Particles(t)$ **do**
    $(\mathbf{V}, \mathbf{W}, prob) \leftarrow$ Sample_typical$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_{t+1}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
    $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot EV(t), 1)\}$
    $Add\_Des \leftarrow$ Sample_additive$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Y}_{t+1}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M)$
    **for** $(\mathbf{V}, \mathbf{W}, prob) \in Add\_Des$ **do**
      $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob \cdot EV(t), 1)\}$
    **end for**
  **end for**
  **for** $k \in \{1, ..., max\_horizon\}$ **do**
    **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in Particles(t-k+1)$ **do**
      $\tilde{\mathbf{Y}} \leftarrow \left[\mathbf{Y}^T_{t-k+2}, ..., \mathbf{Y}^T_{t+1}\right]^T$
      $Inn\_Des \leftarrow$ BS_inn$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\mathbf{Y}}, \mathbf{A}, \mathbf{C}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I, M, k)$
      **for** $(\mathbf{V}, \mathbf{W}, prob) \in Inn\_Des$ **do**
        $Cand \leftarrow Cand \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, \frac{prob \cdot EV(t+1-k)}{max\_horizon}), k)\}$
      **end for**
    **end for**
  **end for**
  $EV(t+1) \leftarrow 0$ ▷ Calculate estimate of evidence at time $t+1$
  **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, k) \in Cand$ **do**
    $EV(t+1) \leftarrow EV(t+1) + prob/|Cand|$
  **end for**
  $Desc \leftarrow$ Resample$(N, Cand)$ ▷ Resample particles
  $Particles(t) \leftarrow \{\}$ ▷ Calculate $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\Sigma}_{t+1}$ for each particle
  **for** $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{V}, \mathbf{W}, prob, k) \in Descendants$ **do**
    $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+2-k}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \mathbf{V}\boldsymbol{\Sigma}_A, \mathbf{W}\boldsymbol{\Sigma}_I)$
    **if** $k > 1$ **then**
      **for** $i \in \{2, ..., k\}$ **do**
        $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow$ KF_Upd$(\mathbf{Y}_{t+1+i-k}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{C}, \mathbf{A}, \boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_I)$
      **end for**
    **end if**
    $Particles(t+1) \leftarrow Particles(t+1) \cup \{(\boldsymbol{\mu}, \boldsymbol{\Sigma})\}$
  **end for**
**end for**