Outlier-robust Kalman Filtering through Generalised Bayes

Gerardo Duran-Martin [1 2] **Matias Altamirano** [3] **Alexander Y. Shestopaloff** [1 4] **Leandro Sánchez-Betancourt** [2 5]
Jeremias Knoblauch [3] **Matt Jones** [6] **François-Xavier Briol** [3] **Kevin Murphy** [7]

Abstract

We derive a novel, provably robust, and closed-form Bayesian update rule for online filtering in state-space models in the presence of outliers and misspecified measurement models. Our method combines generalised Bayesian inference with filtering methods such as the extended and ensemble Kalman filter. We use the former to show robustness and the latter to ensure computational efficiency in the case of nonlinear models. Our method matches or outperforms other robust filtering methods (such as those based on variational Bayes) at a much lower computational cost. We show this empirically on a range of filtering problems with outlier measurements, such as object tracking, state estimation in high-dimensional chaotic systems, and online learning of neural networks.

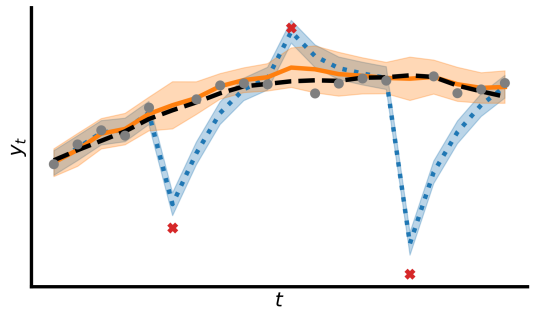

Figure 1. First state component of the SSM (20). The grey dots are measurements sampled from (20) and the red crosses are measurements sampled from an outlier measurement process. The dotted blue line shows the KF posterior mean estimate and the solid orange line shows our proposed WoLF posterior mean estimate. The regions around the posterior mean cover two standard deviations. For comparison, the dashed black line shows the true sampled state process.

1. Introduction

Probabilistic state-space models (SSMs) are widely used to address problems in time-series forecasting, online learning, tracking problems, and signal processing. A key challenge in SSMs is to perform online (sequential) posterior inference, also known as Bayesian filtering. If the model is linear and the and the dynamics are known, then the Kalman filter (KF) algorithm is the optimal filter in terms of minimal mean squared error (Morris, 1976). If the model is not linear, approximate inference methods must be used. Although there are a number of techniques available, in this paper we focus on the extended Kalman filter (EKF) (see e.g., Sarkka & Svensson, 2023) and the ensemble Kalman filter (EnKF) (see e.g., Roth et al., 2017a), because they both admit closed-form Bayesian updates to Gaussian posterior approximations and scale to high dimensions.

In a Bayesian setting, the KF and KF-like methods typically use a Gaussian measurement model for computational convenience. However, a weakness of this assumption is that extreme observations, such as outliers, are not well modelled by Gaussians, leading to model misspecification (Grubbs, 1969; Hampel et al., 2011). Moreover, in practice, the true dynamics are often different from the one assumed by the filter. This could be because the measurements are corrupted, or because the user does not have access to the measurement function. For example, in tracking problems, sensor measurement errors record the wrong position of an object; similarly, in economics and finance, there are instances of data providers sending time series with erroneous values (see e.g. Liu, 2020, for a survey).

There is a large literature that studies the filtering problem in the presence of outliers and misspecified measurement models[1]. Some approaches extend the state space by introducing

[1]School of Mathematical Sciences, Queen Mary University, London, UK [2]Oxford-Man Institute of Quantitative Finance, University of Oxford, UK [3]Department of Statistical Science, University College London, London, United Kingdom [4]Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada [5]Mathematical Institute, University of Oxford, UK [6]Institute for Cognitive Science, University of Colorado Boulder, US [7]Google DeepMind. Correspondence to: Gerardo Duran-Martin <g.duranmartin@qmul.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Alternatively, a measurement model can be referred to as an observation model.

hierarchical priors on the measurement model. Since closed-form estimation of the state is not generally feasible, a factorised variational Bayes (VB) approach can be employed to obtain a fixed-point solution (see e.g., Ting et al., 2007; Agamennoni et al., 2012; Huang et al., 2016; Nurminen et al., 2015; Wang et al., 2018; Piché et al., 2012). Alternatively, Huberised (E)KF-type methods are optimisation-based approaches to filtering, which minimise the Huber loss; see Boncelet & Dickinson (1983); Karlgaard (2015); Das (2023) and the references therein. Recently, Boustati et al. (2020) proposed a particle filter (PF) approach that uses generalised Bayes (GB) to handle model misspecification from outliers; this builds on work showing that GB updates are theoretically sound and robust to outliers and model misspecification (see, e.g., Bissiri et al., 2016; Jewson et al., 2018; Knoblauch et al., 2022; Fong et al., 2021; Jewson & Rossell, 2022; Husain & Knoblauch, 2022; Matsubara et al., 2023; 2022). However, these approaches can be slow, due to the need to perform iterative optimisation per step (for VB) (Knoblauch et al., 2018), or the need to use a large number of samples (for PF) (Boustati et al., 2020). See Appendix A for an overview of a number of these methods.

In this paper, we propose a novel approach that tackles the filtering problem in the presence of outliers and measurement model misspecification. Our method is based on the GB approach where one replaces the log-likelihood of the measurement process with a loss function. We call our method the *weighted observation likelihood filter (WoLF)* because it uses a weighted log-likelihood as loss. A key advantage of this choice is that we can have closed-form (conjugate) update equations that compute an approximate Gaussian posterior. We derive WoLF variants for the KF, the EKF, and the ensemble Kalman Filter (EnKF).

Our approach has several key advantages over prior work: (i) it is fast and has similar computational cost to the KF thanks to closed-form updates, (ii) it is flexible to the form of misspecification, (iii) it is provably robust to outliers, and (iv) it is easy to implement and straightforward to apply to other filtering methods. See Figure 1 for an illustrative example.

The remainder of the paper proceeds as follows: in Section 2 we introduce the filtering problem and some common algorithms to approximate a Gaussian posterior. In Section 3 we present our method and derive WoLF variants to the KF, the EKF, and the EnKF. Next, in Section 3.4 we show that for certain choices of weighting functions, our method is provably robust. Finally, in Section 4, we show empirically that our method matches and in some cases outperforms previous robust filtering methods at a lower running time. We consider a variety of filtering problems, including 2d-tracking, state estimation in high-dimensional chaotic systems, and online learning of neural networks. Our code can be found at https://github.com/gerdm/weighted-likelihood-filter.

2. Background: Filtering in SSMs

We briefly review Kalman filtering and the main extensions we consider in this work. In what follows, $m, d \in \mathbb{N}$ are the dimensions of the state and measurement processes, and $p(\cdot)$ is used for densities. Given an initial state $\boldsymbol{\theta}_0 \in \mathbb{R}^m$, a Markovian state-space model (SSM) is defined as

$$\boldsymbol{\theta}_t = f_t(\boldsymbol{\theta}_{t-1}) + \boldsymbol{\phi}_t, \tag{1}$$

$$\boldsymbol{y}_t = h_t(\boldsymbol{\theta}_t) + \boldsymbol{\varphi}_t, \tag{2}$$

for $t \in \{1, \dots, T\}$. Here, $\boldsymbol{\theta}_t \in \mathbb{R}^m$ is the (latent) state vector, $\boldsymbol{y}_t \in \mathbb{R}^d$ is the (observed) measurement vector, $f_t : \mathbb{R}^m \to \mathbb{R}^m$ is the dynamics function, $h_t : \mathbb{R}^m \to \mathbb{R}^d$ is the measurement function,[2] $\boldsymbol{\phi}_t$ is a zero-mean Gaussian-distributed random vector with known covariance matrix $\mathbf{Q}_t$, and $\boldsymbol{\varphi}_t$ is any zero-mean random vector representing the measurement noise. The meaning of the state depends on the application; for example, it can be the position of an object, the state of the atmosphere, or the weights of a neural network, as we will see in the results section.

The filtering of an SSM consists of two steps: the predict step and the update step. Given the result of a previous update step (i.e., the iterative posterior) $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{y}_{1:t-1})$, the predict step estimates the prior predictive distribution

$$p(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) p(\boldsymbol{\theta}_{t-1}|\boldsymbol{y}_{1:t-1}) \mathrm{d}\boldsymbol{\theta}_{t-1}, \tag{3}$$

and the update step estimates the new posterior distribution

$$p(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) \propto p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}). \tag{4}$$

where $\propto$ indicates equality up to a multiplicative normalisation constant.

2.1. The Kalman filter

Suppose the SSM is linear and Gaussian, that is,

$$\begin{aligned} \boldsymbol{\theta}_t &= \mathbf{F}_t \, \boldsymbol{\theta}_{t-1} + \boldsymbol{\phi}_t, \\ \boldsymbol{y}_t &= \mathbf{H}_t \, \boldsymbol{\theta}_t + \boldsymbol{\varphi}_t, \end{aligned} \tag{5}$$

with $\boldsymbol{\varphi}_t$ a zero-mean Gaussian with known covariance matrix $\mathbf{R}_t$, $\mathbf{F}_t \in \mathbb{R}^{m \times m}$, and $\mathbf{H}_t \in \mathbb{R}^{d \times m}$. Then, given the initial condition $p(\boldsymbol{\theta}_0) = \mathcal{N}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ — a Gaussian density with known mean $\boldsymbol{\mu}_0 \in \mathbb{R}^m$ and covariance $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{m \times m}$ — the exact Bayesian predict and update steps are given by

$$p(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t \,|\, \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}), \tag{6}$$

$$p(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) = \mathcal{N}(\boldsymbol{\theta}_t \,|\, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \tag{7}$$

[2] In many applications, the function h_t is modulated by an exogenous feature vector $\boldsymbol{x}_t$. See e.g., Section 4.2.

with prior predictive mean and covariance given by

$$\begin{aligned}\boldsymbol{\mu}_{t|t-1} &= \mathbf{F}_t\,\boldsymbol{\mu}_{t-1},\\ \boldsymbol{\Sigma}_{t|t-1} &= \mathbf{F}_t\,\boldsymbol{\Sigma}_{t-1}\,\mathbf{F}_t^\intercal + \mathbf{Q}_t,\end{aligned} \tag{8}$$

and posterior mean and covariance given by

$$\begin{aligned}\boldsymbol{\Sigma}_t^{-1} &= \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\intercal\,\mathbf{R}_t^{-1}\,\mathbf{H}_t,\\ \mathbf{K}_t &= \boldsymbol{\Sigma}_t\,\mathbf{H}_t^\intercal\,\mathbf{R}_t^{-1},\\ \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\,(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t),\end{aligned} \tag{9}$$

where $\hat{\boldsymbol{y}}_t = \mathbf{H}_t\,\boldsymbol{\mu}_{t|t-1}$ is the predicted observation and $\mathbf{K}_t$ is the Kalman gain matrix used to map the error (residual) vector in observation space to an update in latent state space.

2.2. The Extended Kalman filter

When the state and measurement functions are non-linear, a common approach is to introduce a Gaussian posterior density $q(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) = \mathcal{N}(\boldsymbol{\theta}_t \mid \boldsymbol{\mu}_t,\, \boldsymbol{\Sigma}_t)$ that approximates the posterior density $p(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t})$ through a Kullback-Leibler projection. This is done using Gaussian approximations of the joint densities $q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\boldsymbol{y}_{1:t-1})$ and $q(\boldsymbol{\theta}_t, \boldsymbol{y}_t|\boldsymbol{y}_{1:t-1})$ and then performing a closed-form Bayesian update; see Section 8.4 in Sarkka & Svensson (2023) for details.

The EKF is a special case of the above in which one linearises f_t in (1) and h_t in (2), and assumes a Gaussian measurement noise $\boldsymbol{\varphi}_t$. As a consequence, the predict and update equations resemble those of the standard KF. More precisely, the EKF replaces the mean of the state transition density $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ in (1) with

$$\mathbb{E}[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}] \approx \mathbf{F}_t(\boldsymbol{\theta}_{t-1} - \boldsymbol{\mu}_{t-1}) + f_t(\boldsymbol{\mu}_{t-1}) =: \bar{\boldsymbol{\mu}}_{t|t-1}, \tag{10}$$

and the measurement mean of $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ in (2) with

$$\mathbb{E}[\boldsymbol{y}_t|\boldsymbol{\theta}_t] \approx \mathbf{H}_t(\boldsymbol{\theta}_t - \boldsymbol{\mu}_{t|t-1}) + h_t(\boldsymbol{\mu}_{t|t-1}) =: \bar{\boldsymbol{y}}_t, \tag{11}$$

where $\boldsymbol{\mu}_{t|t-1} = \mathbb{E}[\bar{\boldsymbol{\mu}}_{t|t-1}|\boldsymbol{\mu}_{t-1}] = f_t(\boldsymbol{\mu}_{t-1})$, $\mathbf{F}_t$ is the Jacobian of f_t evaluated at $\boldsymbol{\mu}_{t-1}$, $\mathbf{H}_t$ is the Jacobian of h_t evaluated at $\boldsymbol{\mu}_{t|t-1}$, and $\hat{\boldsymbol{y}}_t = \mathbb{E}[\bar{\boldsymbol{y}}_t] = h_t(\boldsymbol{\mu}_{t|t-1})$. The linearisation of the transition and measurement around the respective previous means allows the resulting predict and update equations to closely resemble (8) and (9). Therefore, the algorithm remains relatively scalable and more efficient low-rank extensions can also be derived (see e.g. Chang et al., 2023; Lambert et al., 2023; Cartea et al., 2023).

2.3. The ensemble Kalman filter

The ensemble Kalman filter (EnKF) was developed as an alternative to the extended Kalman filter for high-dimensional and highly non-linear state dynamics (see e.g., Evensen, 1994; Burgers et al., 1998; Roth et al., 2017a). This method avoids the need to compute Jacobians and the storage of an explicit $m \times m$ posterior covariance matrix, by instead representing the belief state with an ensemble of $N \in \mathbb{N}$ particles $\hat{\boldsymbol{\theta}}_t^{(i)} \in \mathbb{R}^m$ for $i = 1, \ldots, N$ evolved through time. This enables the method to scale to high-dimensional state spaces with complex nonlinear dynamics, such as those frequently arising in data assimilation for weather forecasting (Evensen, 2009). In the EnKF, the predict step samples (1) to obtain $\hat{\boldsymbol{\theta}}_{t|t-1}^{(i)}$. Then, the update step samples predictions $\hat{\boldsymbol{y}}_{t|t-1}^{(i)} \in \mathbb{R}^d$, for each particle, according to

$$\hat{\boldsymbol{y}}_{t|t-1}^{(i)} \sim \mathcal{N}\left(h_t\left(\hat{\boldsymbol{\theta}}_{t|t-1}^{(i)}\right), \mathbf{R}_t\right). \tag{12}$$

The particles are then updated according to

$$\hat{\boldsymbol{\theta}}_t^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + \bar{\mathbf{K}}_t\big(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\big), \tag{13}$$

with gain matrix $\bar{\mathbf{K}}_t$ calculated from the ensemble. See Appendix D.3 for details.

3. The weighted observation likelihood filter

Our method is based on the GB approach where one modifies the update step in (4) to use a loss function $\ell_t : \mathbb{R}^m \to \mathbb{R}$ in place of the negative log-likelihood of the measurement process. This gives the generalised posterior

$$q(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) \propto \exp(-\ell_t(\boldsymbol{\theta}_t))q(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}). \tag{14}$$

We propose to gain robustness to outliers in observation space by taking the loss function to be the model's negative log-likelihood scaled by a data-dependent weighting term

$$\ell_t(\boldsymbol{\theta}_t) = -W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)\,\log q(\boldsymbol{y}_t|\boldsymbol{\theta}_t), \tag{15}$$

with $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the weighting function and $q(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ the modelled measurement process. We call our method the *weighted observation likelihood filter (WoLF)*. To specify an instance of our method, ones needs to define the likelihood $q(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ and the weighting function W. In the next subsections, we show the flexibility of WoLF and derive weighted-likelihood-based KF, EKF, and EnKF algorithms. Setting $W(\boldsymbol{y}_t, \bar{\boldsymbol{y}}_t) = 1$ trivially recovers existing methods, but we will instead use non-constant weighting functions inspired by the work of (Barp et al., 2019; Matsubara et al., 2022; Altamirano et al., 2023a;b).

3.1. Linear weighted observation likelihood filter

The following proposition gives a closed-form solution for the update step of WoLF under a linear measurement function and a Gaussian likelihood (see Appendix C.1 for the proof).

Proposition 3.1. *Consider the linear-Gaussian SSM* (5) *with weighting function* $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. *Then, the update step of WoLF with loss function* (15) *is given by* (9) *with* $\mathbf{R}_t^{-1}$ *replaced by* $\bar{\mathbf{R}}_t^{-1} = W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)\,\mathbf{R}_t^{-1}$.

The resulting predict and update steps for WoLF under linear dynamics and zero-mean Gaussians for the state and measurement process are shown in Algorithm 1.

Algorithm 1 WoLF predict and update step

Require: $\mathbf{F}_t$, $\mathbf{Q}_t$ // predict step
$\boldsymbol{\mu}_{t|t-1} \leftarrow \mathbf{F}_t\, \boldsymbol{\mu}_{t-1}$
$\boldsymbol{\Sigma}_{t|t-1} \leftarrow \mathbf{F}_t\, \boldsymbol{\Sigma}_{t-1}\, \mathbf{F}_t^\intercal + \mathbf{Q}_t$
Require: $\boldsymbol{y}_t$, $\mathbf{H}_t$, $\mathbf{R}_t$ // update step
$\hat{\boldsymbol{y}}_t \leftarrow \mathbf{H}_t\, \boldsymbol{\mu}_{t|t-1}$
$w_t \leftarrow W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$
$\boldsymbol{\Sigma}_t^{-1} \leftarrow \boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t^2\, \mathbf{H}_t^\intercal\, \mathbf{R}_t^{-1}\, \mathbf{H}_t$
$\mathbf{K}_t \leftarrow w_t^2\, \boldsymbol{\Sigma}_t\, \mathbf{H}_t^\intercal\, \mathbf{R}_t^{-1}$
$\boldsymbol{\mu}_t \leftarrow \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\, (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)$

The computational complexity of WoLF under linear dynamics matches that of the KF, i.e., $O(m^3)$. Alternative robust filtering algorithms require multiple iterations per measurement to achieve robustness and stability, making them significantly slower; see Table 1 for the computational complexity for the methods we consider, and Table 2 and Figure 5 for empirical comparisons.

3.2. Nonlinear weighted observation likelihood filter

Our method readily extends to other nonlinear filtering algorithms. For example, a WoLF version of the EKF is obtained by introducing a weighting function to (11) yielding the approximate log-likelihood

$$\log q(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) \log \mathcal{N}(\boldsymbol{y}_t \,|\, \bar{\boldsymbol{y}}_t,\, \mathbf{R}_t). \tag{16}$$

Similarly, we can derive a weighted ensemble Kalman filter by weighting error terms in the update step (13); see Appendix D.3 for details. Finally, our method can be easily extended to handle measurement processes modelled as an exponential family, such as in classification; see Appendix D.1 for details.

3.3. The choice of weighting function

Weighted likelihoods have a well-established history in Bayesian inference and have demonstrated their efficacy in improving robustness (Grünwald, 2012; Holmes & Walker, 2017; Grünwald & van Ommen, 2017; Miller & Dunson, 2018; Bhattacharya et al., 2019; Alquier & Ridgway, 2020; Dewaskar et al., 2023). In this context, the corresponding posteriors are often referred to as fractional, tempered, or power posteriors. In most existing work, the determination of weights relies on heuristics and the assigned weights remain constant across all data points so that $W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = w \in \mathbb{R}$ for all t. In contrast, we dynamically incorporate information from the most recent observations without incurring additional computational costs by defining the weight as a function of the current observation $\boldsymbol{y}_t$ and its prediction $\hat{\boldsymbol{y}}_t = h_t(\boldsymbol{\mu}_{t|t-1})$, which is based on all of the past observations.

To define the weighting function, we take inspiration from previous work for dealing with outliers. In particular, Wang et al. (2018) proposed classifying robust filtering algorithms into two main types: *compensation-based* algorithms, which incorporate information from tail events into the model in a robust way (see, e.g., Huang et al., 2016; Agamennoni et al., 2012), and *detect-and-reject* algorithms, which assume that outlier observations bear no useful information and thus are ignored (see, e.g., Wang et al., 2018; Mu & Yuen, 2015). Below we show how both of these strategies can be implemented using our WoLF method by merely changing the weighting function.

Inverse multi-quadratic weighting function: As an example of a compensation-based method, we follow Altamirano et al. (2023b) and use the Inverse Multi-Quadratic (IMQ) weighting, which in our SSM setting is

$$W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \left(1 + \frac{||\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t||_2^2}{c^2}\right)^{-1/2}, \tag{17}$$

where $c > 0$ is the soft threshold and $\|\cdot\|$ denotes the l_2 norm. We call WoLF with IMQ weighting "WoLF-IMQ".

Mahalanobis-based weighting function: The l_2 norm in the IMQ can be modified to account for the covariance structure of the measurement process by replacing it with the Mahalanobis distance between $\boldsymbol{y}_t$ and $\hat{\boldsymbol{y}}_t$:

$$W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \left(1 + \frac{\|\mathbf{R}_t^{-1/2}(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\|_2^2}{c^2}\right)^{-1/2}. \tag{18}$$

We call WoLF with this weighting function the WoLF-MD method. This type of weighted IMQ function has been used extensively in the kernel literature (see e.g. Chen et al., 2019; Detommaso et al., 2018; Riabiz et al., 2022).

Threshold Mahalanobis-based weighting function: As an example of a detect-and-reject method, we consider

$$W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \begin{cases} 1 & \text{if } \|\mathbf{R}_t^{-1/2}(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\|_2^2 \leq c \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

with $c > 0$ the fixed threshold. The weighting function (19) corresponds to ignoring information from estimated measurements whose Mahalanobis distance to the true measurement is larger than some predefined threshold c. In the linear setting, this weighting function is related to the benchmark method employed in Ting et al. (2007). We refer to WoLF with this weighting function as "WoLF-TMD".

For $\boldsymbol{y}_t \in \mathbb{R}^d$ with $d = m \gg 1$ and diagonal measurement covariance $\mathbf{R}_t = \operatorname{diag}(r_{t,1}, \ldots, r_{t,d})$, the WoLF-TMD function can be modified to weight individual observations so that $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. See Section 4.3 for an example and Appendix D.2 for a discussion.

The proposed weighting functions — the IMQ, the MD, and the TMD — are defined such that $W : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ and therefore can only down-weight observations. This means that our updates are always conservative, i.e., our posteriors will be wider in the presence of outliers (see Figure 1 for an example).

3.4. Theoretical properties

In this section, we prove the outlier-robustness for WoLF-type methods. We use the classical framework of Huber (1981). Consider measurements $\boldsymbol{y}_{1:t}$. We measure the influence of a contamination $\boldsymbol{y}_t^c$ by examining the divergence between the posterior with the original observation $\boldsymbol{y}_t$ and the posterior with the contamination $\boldsymbol{y}_t^c$, which is allowed to be arbitrarily large. As a function of $\boldsymbol{y}_t^c$, this divergence is called the *posterior influence function* (PIF) and was studied in Ghosh et al. (2016); Matsubara et al. (2022); Altamirano et al. (2023a;b). Following Altamirano et al. (2023b), we consider the Kullback-Leibler (KL) divergence, which allows us to obtain closed-form expressions for Gaussians. The PIF is given by

$$\operatorname{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}) = \operatorname{KL}\left(p(\boldsymbol{\theta}_t | \boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1}) \| p(\boldsymbol{\theta}_t | \boldsymbol{y}_t, \boldsymbol{y}_{1:t-1})\right).$$

If $\sup_{\boldsymbol{y}_t^c \in \mathbb{R}^d} |\operatorname{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t})| < \infty$, then the posterior is called outlier-robust, which indicates that as $\|\boldsymbol{y}_t - \boldsymbol{y}_t^c\|_2 \to \infty$, the contamination's effect on the posterior is bounded. This is the Bayesian equivalent to bias-robustness in frequentist statistics.

Theorem 3.2. *Consider the linear Gaussian SSM (5) or its linearised (EKF) approximation. The standard (E)KF posterior has an unbounded PIF and is not outlier robust.*

In contrast, the generalised posterior presented in Proposition 3.1 has bounded PIF and is, therefore, outlier robust for any weighting function W such that $\sup_{\boldsymbol{y}_t \in \mathbb{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) < \infty$ and $\sup_{\boldsymbol{y}_t \in \mathbb{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)^2 \|\boldsymbol{y}_t\|_2 < \infty$.

The proof is in Appendix C.2. In particular, the conditions are satisfied when W is (17), (18), or (19), which are the focus of our paper. Figure 2 shows an empirical validation of this proposition for the 2D tracking problem detailed in Section 4.1. Here, it is clear that the PIF for the standard KF is unbounded, whereas our methods exhibit a bounded influence. In Appendix C.3 we show a version of the theorem above for the EnKF.

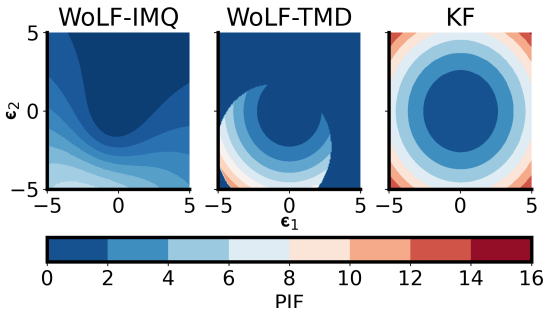

Figure 2. PIF for the 2d tracking problem of Section 4.1. The last measurement $\boldsymbol{y}_t$ is replaced with $\boldsymbol{y}_t^c = \boldsymbol{y}_t + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in [-5, 5] \times [-5, 5]$. We observe that the PIF is asymetric for the weighted methods; this is because the weighting term is a function of the prior predictive and the measurement at time t. See Appendix E.1 for a more detailed explanation.

4. Experiments

In this section, we study the performance of the WoLF methods in multiple filtering settings. Each experiment employs a dataset (or samples data from an SSM), a collection of benchmark methods, and a metric to compare the methods.

For our robust baselines, we make use of three methods that are representative of recent state-of-the-art approaches to robust filtering: the Bernoulli KF of Wang et al. (2018) (**KF-B**), which is an example of a detect-and-reject strategy; the inverse-Wishart filter of Agamennoni et al. (2012) (**KF-IW**), which is an example of a compensation-based strategy; and the Huberised EnKF of Roh et al. (2013) (**Hub-EnKF**), which is an example of a Huberised algorithm. The **KF-B** and **KF-IW** are deterministic and optimise a VB objective to compute a Gaussian approximation to the state posterior (see Appendix A for details). We do not compare against sophisticated hierarchical methods nor methods based on particle filtering, because these do not scale well to high-dimensional state spaces. For the neural network fitting problem, we also consider a variant of online gradient descent (**OGD**) based on Adam (Kingma & Ba, 2017), which uses multiple inner iterations per step (measurement). This method does scale to high-dimensional state spaces, but sadly only gives a maximum a posteriori (MAP) estimate and is not as sample efficient as a robust Bayesian filter.

For experiments where KF or EKF is used as the baseline, we consider the following WoLF variants: (i) the WoLF version with inverse multi-quadratic weighting function (**WoLF-IMQ**), (ii) the thresholded WoLF with Mahalanobis-based weighting function (**WoLF-TMD**). When using the EKF variants, we linearise the state mean (10) and measurement mean (11). For experiments where the ensemble KF (**EnKF**) is taken as the baseline algorithm, we benchmark the performance of the weighted likelihood EnKF with (i)

weighting with averaged-particles (**AP-EnKF**) and (ii) the per-particle weightings (**PP-EnKF**). See Appendix D.3 for a detailed description of the robust EnKF methods.

Method	Cost	#HP	Ref
KF	$O(m^3)$	0	Kalman
KF-B	$O(I\,m^3)$	3	Wang et al.
KF-IW	$O(I\,m^3)$	2	Agamennoni et al.
OGD	$O(I\,m^2)$	2	Bencomo et al.
WoLF-IMQ	$O(m^3)$	1	(Ours)
WoLF-TMD	$O(m^3)$	1	(Ours)

Table 1. Computational complexity of the update step, assuming $d \leq m$ and assuming linear dynamics. Here, I is the number of inner iterations, #HP refers to the number of hyperparameters we tune, and "Cost" refers to the computational complexity.

In each experiment, and unless otherwise specified, we run 100 trials to evaluate each method. The hyperparameters of each method are chosen on the first trial using the Bayesian optimisation (BO) package of Nogueira (2014). BO is a popular derivative-free approach to function maximisation (see e.g. Frazier, 2018). Specifically, we optimise the hyperparameters that minimise the chosen metric on the first run of each experiment. Where a multi-output metric is specified, the minimisation is taken over the maximum of the output. The hyperparameters for KF/EKF-like methods are: the noise scaling and number of inner iterations for the **KF-IW**, the two shape parameters and the number of inner iterations for the **KF-B**, the learning rate and number of inner iterations for the **OGD**, the thresholding value for the **WoLF-TMD**, and the soft threshold hyperparameter for the **WoLF-IMQ**. See Table 1 for a summary.

The results we obtain can be summarised as follows: WoLF-based methods either outperform or match the performance of their counterparts in the metrics we specify below, but typically at a much lower running time.

4.1. Robust KF for tracking a 2D object

We consider the classical problem of estimating the position of an object moving in 2D with constant velocity, which is commonly used to benchmark tracking problems (see e.g., Example 8.2.1.1 in Murphy (2023) or Example 4.5 in Sarkka & Svensson (2023)). The SSM takes the form

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) &= \mathcal{N}(\boldsymbol{\theta}_t \,|\, \mathbf{F}_t\boldsymbol{\theta}_{t-1},\, \mathbf{Q}_t), \\ p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) &= \mathcal{N}(\boldsymbol{y}_t \,|\, \mathbf{H}_t\boldsymbol{\theta}_t,\, \mathbf{R}_t), \end{aligned} \tag{20}$$

where $\mathbf{Q}_t = q\,\mathbf{I}_4$, $\mathbf{R}_t = r\,\mathbf{I}_2$, $(\boldsymbol{\theta}_{0,t}, \boldsymbol{\theta}_{1,t})$ is the position, $(\boldsymbol{\theta}_{2,t}, \boldsymbol{\theta}_{3,t})$ is the velocity,

$$\mathbf{F}_t = \begin{pmatrix} 1 & 0 & \Delta & 0 \\ 0 & 1 & 0 & \Delta \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{H}_t = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

$\Delta = 0.1$ is the sampling rate, $q = 0.10$ is the system noise, $r = 10$ is the measurement noise, and $\mathbf{I}_K$ is a $K \times K$ identity matrix. We simulate 500 trials, each with 1,000 steps. For each method, we compute the scaled RMSE metric $J_{T,i} = \sqrt{\sum_{t=1}^{T}(\boldsymbol{\theta}_{t,i} - \boldsymbol{\mu}_{t,i})^2}$ for $i \in \{0, 1, 2, 3\}$ as well as the total running time (relative to the **KF**).

In our experiments, the true data generating process is one of two variants of (20). The first variant (which we call **Student observations**) corresponds to a system whose measurement process comes from the Student-t likelihood:

$$\begin{aligned} p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) &= \mathrm{St}(\boldsymbol{y}_t \,|\, \mathbf{H}_t\boldsymbol{\theta}_t,\, \mathbf{R}_t, \nu_t) \\ &= \int_0^\infty \mathcal{N}\left(\boldsymbol{y}_t \,|\, \mathbf{H}_t\boldsymbol{\theta}_t, \frac{\mathbf{R}_t}{\tau}\right) \mathrm{Gam}\left(\tau | \frac{\nu_t}{2}, \frac{\nu_t}{2}\right) d\tau, \end{aligned} \tag{21}$$

with $\mathrm{Gam}(\cdot|a, b)$ the density of a Gamma distribution with shape a and rate b, and $\nu_t = 2.01$. The second variant (which we call **mixture observations**) corresponds to a system where the mean of the observations changes sporadically. Instances of this variant can occur as a form of human error or a software bug in a data-entry program. To emulate this scenario, we modify (20) by using the following mixture model for the observation process:

$$\begin{aligned} p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) &= \mathcal{N}(\boldsymbol{y}_t \,|\, \boldsymbol{m}_t,\, \mathbf{R}_t), \\ \boldsymbol{m}_t &= \begin{cases} \mathbf{H}_t\,\boldsymbol{\theta}_t & \text{w.p. } 1 - p_\epsilon, \\ 2\,\mathbf{H}_t\,\boldsymbol{\theta}_t & \text{w.p. } p_\epsilon, \end{cases} \end{aligned} \tag{22}$$

where $p_\epsilon = 0.05$.

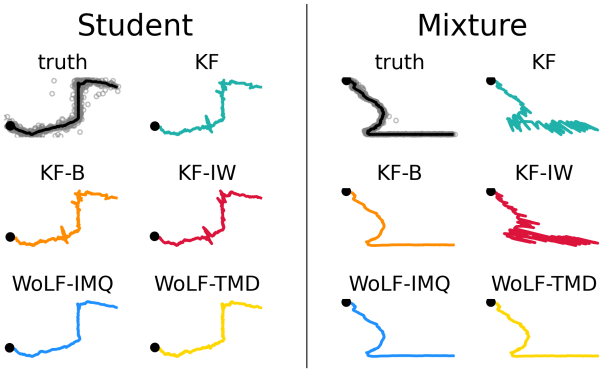

Figure 3. The left panel shows a sample path using the Student variant and the right panel shows a sample path using the mixture variant. The top left figure on each panel shows the true underlying state in black, and the measurements as grey dots.

Results Figure 3 shows a sample of each variant along with the filtered state for each method. For the Student variant (left panel), the **WoLF-IMQ** and the **WoLF-TMD** estimate the true state more closely than the competing methods. Both the **KF-IW** and the **KF-B** look comparable

to the KF, which are not robust to outliers. For the mixture variant (right panel),[3] we see that the WoLF-IMQ, the WoLF-TMD, and the KF-B filter the true state correctly. In contrast, the KF-IW and the KF are not robust to outliers.[4]

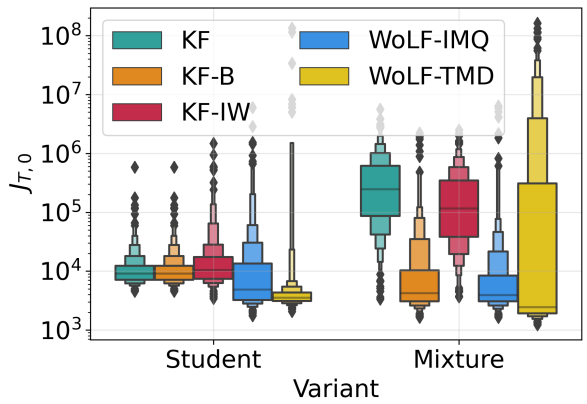

Figure 4. Distribution (across 500 2d tracking trials) of RMSE for first component of the state vector, $J_{T,0}$. Left panel: Student observation model. Right panel: Mixture observation model.

The results in Figure 3 hold for multiple trials as shown in Figure 4, which plots the distribution of the errors in the first component of the state vector. The heavy tail of the WoLF-TMD on the mixture observations variant, and the distribution of $J_{T,k}$ for all k are studied in Appendix E.2.1.

Method	Student	Mixture
KF-B	2.0x	3.7x
KF-IW	1.2x	5.3x
WoLF-IMQ (ours)	1.0x	1.0x
WoLF-TMD (ours)	1.0x	1.0x

Table 2. Mean slowdown rate over KF.

Table 2 shows the median slowdown (in running time) to process the measurements relative to the KF. The slowdown for method X is obtained diving the running time of method X over the running time of the KF. Under the Student variant, the WoLF-IMQ, the WoLF-TMD, and the KF-IW have similar running time to the KF. In contrast, the KF-B takes twice the amount of time. Under the mixture variant, the KF-B and the KF-IW are almost four times and five times slower than the KF respectively. The changes in slowdown rate are due the number of inner iterations that were chosen during the first trial.

4.2. Robust EKF for online MLP regression (UCI)

In this section, we benchmark the methods using a corrupted version of the tabular UCI regression datasets.[5] Similar to other papers that deal with non-linear online learning (see, e.g. Chang et al., 2023), we consider a single-hidden-layer multi-layered perceptron (MLP) with twenty hidden units and a real-valued output unit. In this experiment, the state dimension (number of parameters in the MLP) is $m = (n_{\text{in}} \times 20 + 20) + (20 \times 1 + 1)$, where n_{in} is the dimension of the feature $\boldsymbol{x}_t$.[6] One of the main advantages of using a Bayesian filtering method for fitting neural networks (compared to using OGD) is the ability to handle non-stationary distributions (see e.g. Chang et al., 2023). Below, we take a static state (see e.g. Lambert et al., 2021), so that the prior predictive mean is $\boldsymbol{\mu}_{t|t-1} = \boldsymbol{\mu}_{t-1}$. In Appendix E.3 we study online learning with non-stationary environments.

Each trial is carried out as follows: first, we randomly shuffle the rows in the dataset; second, we divide the dataset into a warmup dataset (10% of rows) and a corrupted dataset (remaining 90% of rows); third, we normalise the corrupted dataset using min-max normalisation from the warmup dataset; fourth, with probability $p_\epsilon = 0.1$, we replace a measurement $\boldsymbol{y}_t \in \mathbb{R}$ with a corrupted data point $\boldsymbol{u}_t \sim \mathcal{U}[-50, 50]$; and fifth, we run each method on the corrupted dataset.

For each dataset and for each method, we evaluate the prior predictive RMedSE $= \sqrt{\text{median}\{(\boldsymbol{y}_t - h_t(\boldsymbol{\mu}_{t|t-1}))^2\}_{t=1}^T}$, which is the squared root of the median squared error between the measurement $\boldsymbol{y}_t$ and the prior predictive $h_t(\boldsymbol{\mu}_{t|t-1}) = h(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{x}_t)$.[7] Here, h is the MLP. We also evaluate the average time step of each method, i.e., we run each method and divide the total running time by the number of samples in the corrupted dataset.

Results Figure 5 shows the percentage change of the RMedSE and the percentage change of running time with respect to those of the OGD for all corrupted UCI datsets. Given the computational complexity of the remaining methods, ideally, a robust Bayesian alternative to the OGD should be as much to the left as possible on the x-axis (rel. time step) and as low as possible on the y-axis (rel. RMedSE). We observe that the WoLF-IMQ and the WoLF-TMD have both of these traits. In particular, we observe that the only two points in the third quadrant are those of the WoLF-IMQ

[3] The top left figure in the right panel is cropped, see Figure 13 for the uncropped version.

[4] KF-B removes outliers that bear no information according to some criterion, but in the Student-t case, it fails. KF-IW, on the other hand, estimates a measurement covariance rather than the dispersion of such a measurement covariance. In this sense, it is misspecified in both cases.

[5] The dataset is available at https://github.com/yaringal/DropoutUncertaintyExps.

[6] See Table 3 for the values that n_{in} takes for each dataset.

[7] We use median instead of mean because we have outliers in measurement space.

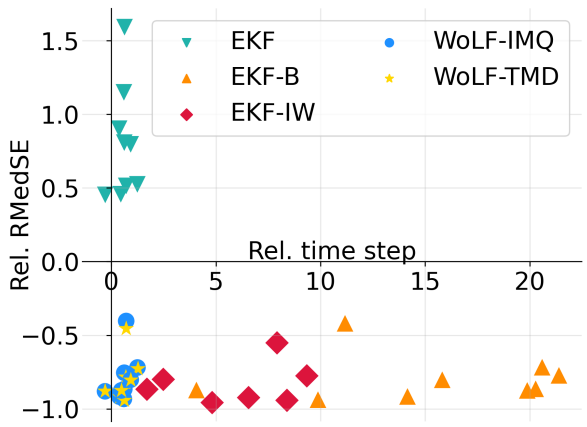

Figure 5. RMedSE versus time per step (relative to the OGD minus 1) across the corrupted UCI datasets.

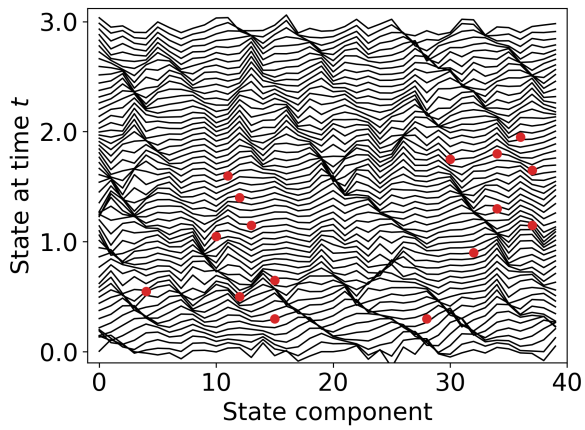

Figure 6. Sample of the Lorenz process with $d = 40$. Here, the "waves" move westward. The red dots represent where measurement outliers occur. We take $p_\epsilon = 0.01$ and $\Delta = 0.05$ integrated over $T = 60$ steps.

and the **WoLF-TMD**. Note that the **EKF-IW** and the **EKF-B** have much higher relative running time and the **EKF** has much higher relative RMedSE.

4.3. Robust EnKF for Lorenz96 model

We consider a modified version of the Lorenz96 model that is commonly used to simulate the atmosphere (see e.g. Lorenz, 2006; Arnold et al., 2013)

For a fixed $\Delta > 0$, the SSM is given by

$$\frac{\boldsymbol{\theta}_{t+\Delta,i} - \boldsymbol{\theta}_{t,i}}{\Delta} = \Big(\boldsymbol{\theta}_{t,i+1} - \boldsymbol{\theta}_{t,i-2}\Big)\boldsymbol{\theta}_{t,i-1} - \boldsymbol{\theta}_{t,i} + \boldsymbol{\phi}_{t,i},$$
$$\boldsymbol{y}_{t,i} = \begin{cases} \boldsymbol{\theta}_{t,i} + \boldsymbol{\varphi}_{t,i} & \text{w.p. } 1 - p_\epsilon, \\ 100 & \text{w.p. } p_\epsilon. \end{cases} \tag{23}$$

Here, $\boldsymbol{\theta}_{t,k}$ is the value of the state component k at step t, $\boldsymbol{\phi}_{t,i} \sim \mathcal{N}(8, 1)$, $\boldsymbol{\varphi}_{t,i} \sim \mathcal{N}(0, 1)$, $p_\epsilon = 0.001$, $i = 1, \ldots, d$, $t = 1, \ldots, T$, with $T \gg 1$ the number of steps, and we use the convention $\boldsymbol{\theta}_{t,d+k} = \boldsymbol{\theta}_{t,k}$, $\boldsymbol{\theta}_{t,-k} = \boldsymbol{\theta}_{t,d-k}$. Similar to Roth et al. (2017a), we integrate the state process in (23) to match the formulation in (1) using the Runge-Kutta-4 (RK4) procedure with discretisation step $\Delta = 0.05$, integrated over $T = 10^3$ steps, $N = 1,000$ number of particles, and $d = m = 100$. A run of the state process is shown in Figure 6. Note that the probability of an outlier happening on any state component at any timestep is $p_\epsilon \times d = 0.1$.

In this experiment, **EnKF** is the baseline. As in Roth et al. (2017b), we use the metric $L_t = \sqrt{\frac{1}{d}(\boldsymbol{\theta}_t - \boldsymbol{\mu}_t)^\intercal(\boldsymbol{\theta}_t - \boldsymbol{\mu}_t)}$ to measure the in-state RMSE.

Results An evaluation of L_t for the **EnKF**, the **AP-EnKF**, the **PP-EnKF**, and the **Hub-EnKF** is shown in the top row of Figure 7. The grey vertical lines denote timesteps where an outlier event happened, i.e., at least one entry of $\boldsymbol{y}_t$ is 100.

The top plot shows that the **Hub-EnKF** and **AP-EnKF** have an almost-identical behaviour. However, the bottom row of Figure 7 shows that the **AP-EnKF** and the **PP-EnKF** are more robust to the choice of threshold c compared to the **Hub-EnKF**. This is because the **Hub-EnKF** makes updates with outlier observations clipped at c, whereas WoLF-like methods disregard error measurements above c. We present the results of the EnKF when the number of particles is less than the number of state components in Appendix E.2.3. In the experiment, we modify the algorithms to make use of the covariance inflation correction first proposed by Anderson (1999). The conclusions from this section extend to the covariance inflation case.

5. Conclusion

We introduced a provably robust filtering algorithm based on generalised Bayes which we call the weighted observation likelihood filter or WoLF. Our algorithm is as fast as the KF, has closed-form update equations, and is straightforward to apply to various filtering methods. The superior performance of the WoLF is shown on a wide range of filtering problems. In contrast, alternative robust methods either have higher computational complexity than the WoLF, or similar computational complexity but not higher performance.

Future work will investigate how to overcome the limitations of our approach. For example, (i) not being robust to outliers in the state-process, (ii) the assumption of a known covariance $\mathbf{R}_t$, (iii) the assumption of known dynamics for the covariance $\mathbf{Q}_t$, and (iv) the assumption of a unimodal posterior.

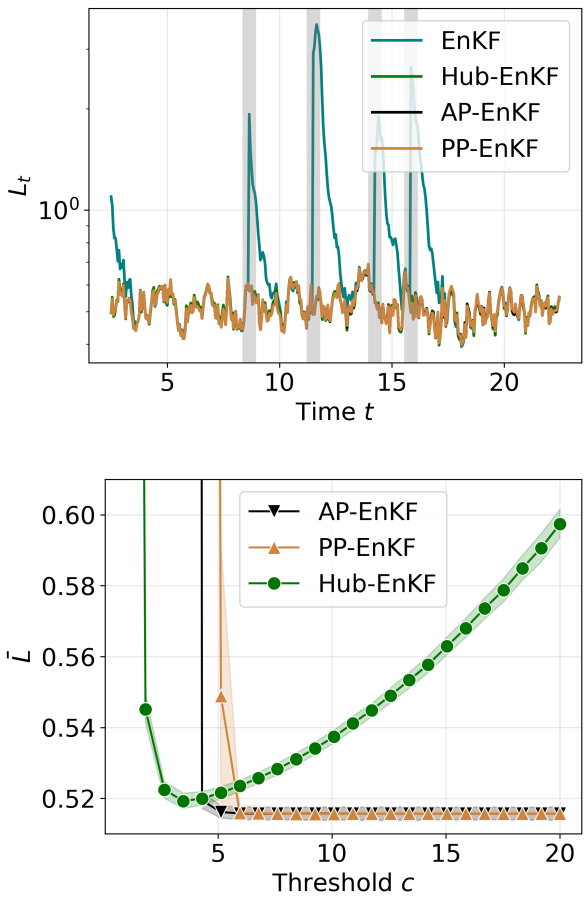

Figure 7. The top row shows a sample run of the **EnKF**, the **AP-EnKF**, the **PP-EnKF**, and the **Hub-EnKF**; outlier events are shown in grey vertical bars. The bottom row shows the bootstrap estimate of L_T over 20 runs and 500 bootstrapped samples as a function of the c hyperparameter.

Acknowledgements

JK and FXB were supported by the EPSRC grant [EP/Y011805/1].

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agamennoni, G., Nieto, J. I., and Nebot, E. M. Approximate inference in state-space models with heavy-tailed noise. *IEEE Transactions on Signal Processing*, 2012.

Agee, W. S., Dunn, B. A., and NM, W. S. M. R. Robust filtering and smoothing via gaussian mixtures. *NTIS, SPRINGFIELD, VA, 1980, 34*, 1980.

Alquier, P. and Ridgway, J. Concentration of tempered posteriors and of their variational approximations. 2020.

Altamirano, M., Briol, F.-X., and Knoblauch, J. Robust and scalable Bayesian online changepoint detection. In *International Conference on Machine Learning*, 2023a.

Altamirano, M., Briol, F.-X., and Knoblauch, J. Robust and conjugate Gaussian process regression. *arXiv:2311.00463*, 2023b.

Anderson, J. L. A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly weather review*, 127(12):2741–2758, 1999.

Arnold, H., Moroz, I., and Palmer, T. Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991): 20110479, 2013.

Arroyo, A., Cartea, A., Moreno-Pino, F., and Zohren, S. Deep attentive survival analysis in limit order books: Estimating fill probabilities with convolutional-transformers. *Quantitative Finance*, pp. 1–23, 2024.

Barp, A., Briol, F.-X., Duncan, A. B., Girolami, M., and Mackey, L. Minimum Stein discrepancy estimators. In *Neural Information Processing Systems*, pp. 12964–12976, 2019.

Bencomo, G. M., Snell, J. C., and Griffiths, T. L. Implicit maximum a posteriori filtering via adaptive optimization. *arXiv:2311.10580*, 2023.

Bhattacharya, A., Pati, D., and Yang, Y. Bayesian fractional posteriors. 2019.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

Boncelet, C. G. and Dickinson, B. W. An approach to robust Kalman filtering. In *IEEE Conference on Decision and Control*, pp. 304–305, 1983.

Boustati, A., Akyildiz, O. D., Damoulas, T., and Johansen, A. Generalised Bayesian filtering via sequential monte carlo. *Advances in neural information processing systems*, 33:418–429, 2020.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Burgers, G., Van Leeuwen, P. J., and Evensen, G. Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, 126(6):1719–1724, 1998.

Calvet, L. E., Czellar, V., and Ronchetti, E. Robust filtering. *Journal of the American Statistical Association*, 110(512): 1591–1606, 2015.

Cao, W., Liu, C., Lan, Z., Li, S. E., Pan, W., and Alessandri, A. Robust bayesian inference for moving horizon estimation. *arXiv preprint arXiv:2210.02166*, 2022.

Cartea, Á., Duran-Martin, G., and Sánchez-Betancourt, L. Detecting toxic flow. *arXiv:2312.05827*, 2023.

Chang, P. G., Duran-Martin, G., Shestopaloff, A. Y., Jones, M., and Murphy, K. Low-rank extended Kalman filtering for online learning of neural networks from streaming data. In *Conference on Lifelong Learning Agents*, 2023.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. Stein point markov chain Monte Carlo. In *International Conference on Machine Learning*, pp. 1011–1021, 2019.

Das, S. *Robust state estimation methods for robotics applications*. PhD thesis, West Virginia University, 2023.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., and Scheichl, R. A Stein variational newton method. *Advances in Neural Information Processing Systems*, 2018.

Dewaskar, M., Tosh, C., Knoblauch, J., and Dunson, D. B. Robustifying likelihoods by optimistically re-weighting data. *arXiv:2303.10525*, 2023.

Duran-Martin, G., Kara, A., and Murphy, K. Efficient online Bayesian inference for neural bandits. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Evensen, G. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

Evensen, G. *Data Assimilation: The Ensemble Kalman Filter*. Springer, 2nd ed. 2009 edition edition, 2009.

Fong, E., Holmes, C., and Walker, S. G. Martingale posterior distributions. *arXiv:2103.15671*, 2021.

Frazier, P. I. A tutorial on Bayesian optimization. *arXiv:1807.02811*, 2018.

Ghosh, S., Delle Fave, F. M., and Yedidia, J. Assumed density filtering methods for learning Bayesian neural networks. In *Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2016.

Grubbs, F. E. Procedures for detecting outlying observations in samples. *Technometrics*, pp. 1–21, 1969.

Grünwald, P. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pp. 169–183, 2012.

Grünwald, P. and van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069 – 1103, 2017.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

Holmes, C. C. and Walker, S. G. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Huang, Y., Zhang, Y., Li, N., and Chambers, J. A robust Gaussian approximate filter for nonlinear systems with heavy tailed measurement noises. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4209–4213, 2016.

Huber, P. J. Robust statistics. *Wiley Series in Probability and Mathematical Statistics*, 1981.

Husain, H. and Knoblauch, J. Adversarial interpretation of Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 553–572. PMLR, 2022.

Jewson, J. and Rossell, D. General Bayesian loss function selection and the use of improper models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. 1640–1665, 2022.

Jewson, J., Smith, J. Q., and Holmes, C. Principled Bayesian minimum divergence inference. *Entropy*, pp. 442, 2018.

Kalman, R. E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(Series D):35–45, 1960.

Karlgaard, C. D. Nonlinear regression Huber–Kalman filtering and fixed-interval smoothing. *Journal of guidance, control, and dynamics*, 38(2):322–330, 2015.

Khan, M. E. and Rue, H. The bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023. URL http://jmlr.org/papers/v24/22-0291.html.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Knoblauch, J., Jewson, J. E., and Damoulas, T. Doubly robust bayesian inference for non-stationary streaming data with $beta$-divergences. *Advances in Neural Information Processing Systems*, 31, 2018.

Knoblauch, J., Jewson, J., and Damoulas, T. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Lambert, M., Bonnabel, S., and Bach, F. The recursive variational Gaussian approximation (R-VGA). *Statistics and Computing.*, 32(1):10, 2021.

Lambert, M., Bonnabel, S., and Bach, F. The limited-memory recursive variational Gaussian approximation (L-RVGA). *Statistics and Computing*, 33(70), 2023. doi: 10.1007/s11222-023-10239-x. URL https://inria.hal.science/hal-03501920.

Lin, W., Khan, M. E., and Schmidt, M. Fast and simple Natural-Gradient variational inference with mixture of exponential-family approximations. In *ICML*, June 2019. URL http://arxiv.org/abs/1906.02914.

Liu, G. Data quality problems troubling business and financial researchers: A literature review and synthetic analysis. *Journal of Business & Finance Librarianship*, 25(3-4):315–371, 2020.

Lorenz, E. N. *Predictability – a problem partly solved*, pp. 40–58. Cambridge University Press, 2006.

Masreliez, C. Approximate non-gaussian filtering with linear state and observation relations. *IEEE Transactions on Automatic Control*, 20(1):107–110, 1975.

Masreliez, C. and Martin, R. Robust bayesian estimation for the linear model and robustifying the kalman filter. *IEEE transactions on Automatic Control*, 22(3):361–371, 1977.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society: Series B*, 84(3):997–1022, 2022.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. Generalized Bayesian inference for discrete intractable likelihood. *Journal of the American Statistical Association*, pp. 1–11, 2023.

Meyr, H. and Spies, G. The structure and performance of estimators for real-time estimation of randomly varying time delay. *IEEE transactions on acoustics, speech, and signal processing*, 32(1):81–94, 1984.

Miller, J. W. and Dunson, D. B. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Morris, J. The Kalman filter: A robust estimator for some classes of linear quadratic problems. *IEEE Transactions on Information Theory*, 22(5):526–534, 1976.

Mu, H.-Q. and Yuen, K.-V. Novel outlier-resistant extended Kalman filter for robust online structural identification. *Journal of Engineering Mechanics*, 141(1):04014100, 2015.

Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

Nogueira, F. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014. URL https://github.com/fmfn/BayesianOptimization.

Nurminen, H., Ardeshiri, T., Piché, R., and Gustafsson, F. Robust inference for state-space models with skewed measurement noise. *IEEE Signal Processing Letters*, 22 (11):1898–1902, 2015.

Ollivier, Y. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2):2930–2961, 2018.

Piché, R., Särkkä, S., and Hartikainen, J. Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate student-t distribution. In *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2012.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. Optimal thinning of MCMC output. *arXiv:2005.03952*, 2022.

Roh, S., Genton, M. G., Jun, M., Szunyogh, I., and Hoteit, I. Observation quality control with a robust ensemble Kalman filter. *Monthly Weather Review*, 141(12):4414–4428, 2013.

Roth, M., Hendeby, G., Fritsche, C., and Gustafsson, F. The ensemble Kalman filter: a signal processing perspective. *Journal on Advances in Signal Processing*, 2017:1–16, 2017a.

Roth, M., Hendeby, G., Fritsche, C., and Gustafsson, F. The ensemble Kalman filter: a signal processing perspective. *EURASIP Journal on Advances in Signal Processing*, pp. 1–16, 2017b.

Sarkka, S. and Svensson, L. *Bayesian Filtering and Smoothing (2nd edition)*. Cambridge University Press, 2023.

Schick, I. C. and Mitter, S. K. Robust recursive estimation in the presence of heavy-tailed observation noise. *The annals of statistics*, 22(2):1045–1080, 1994.

Ting, J.-A., Theodorou, E., and Schaal, S. Learning an outlier-robust kalman filter. In *European Conference on Machine Learning*, pp. 748–756. Springer, 2007.

Wang, H., Li, H., Fang, J., and Wang, H. Robust Gaussian Kalman filter with outlier detection. *IEEE Signal Processing Letters*, 25(8):1236–1240, 2018.

West, M. Robust sequential approximate bayesian estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 43(2):157–166, 1981.

Supplementary Materials

The Appendix is structured as follows. Appendix A provides an overview of outlier-robust filtering methods. Appendix B derives the WoLF-MD method from a maximum-a-posteriori (MAP) estimate perspective. Appendix C collects proofs; more precisely, Appendix C.1 proves Proposition 3.1 and Appendix C.2 proves Theorem 3.2. We also show robustness for the IMQ, MD, and TMD weighting functions. Next, in Appendix D, we discuss the exponential-family, multi-output weighting function, and EnKF extensions to the WoLF. Finally, in Appendix E, we investigate the 2d tracking problem of Section 4.1 in more detail; we conduct robustness checks and we introduce an additional experiment for corrupted non-linear non-stationary learning.

A. Background on existing robust filters and related methods

Robust Bayesian filtering and the minimum variance estimator dates back to Masreliez (1975) and Masreliez & Martin (1977). These methods propose a modified KF-recursion for the linear SSM written in terms of the score function of the measurement prior predictive. In these early works, inference relies on Monte Carlo.

West (1981) follows these earlier works and proposes KF-like updates for non-normal measurement models. That paper also studies whether several popular likelihood functions are robust in the sense of "ignoring outliers" — the analysis includes the Student-t, power exponential, Huber, logistic, and stable-law likelihoods. For the case of Student-t likelihood with one degree of freedom (i.e., Cauchy) and linear dynamics, the update equation for the posterior mean derived in West (1981) is equivalent to ours when using IMQ weights. However, the approach taken in West (1981) cannot recover our TMD scheme, due to being tied to a given choice of measurement model. Furthermore, West (1981) does not provide a theoretical foundation to use their approach in non-linear measurement models.

The work in Meyr & Spies (1984) proposes a scheme to eliminate observations that a KF procedure labels as outliers. Their methodology relies on a "secondary decision system" which checks for discrepancies between the predicted mean and the observation, eliminating observations with high discrepancies. This scheme is analogous to the TMD scheme, however, it is not shown to be provably robust.

Another alternative to robustify measurement models against outliers is the work in Agee et al. (1980). Their paper introduces Gaussian-mixture models for robust filtering and smoothing. Their inference method is based on particle filtering.

To the best of our knowledge, the first work that proposes robust filtering in the context of robust statistics is Calvet et al. (2015). Their robust filter follows Masreliez & Martin (1977) and is based on a "Huberisation" of the derivative of the log-measurement density (score function), which then they integrate. See Schick & Mitter (1994) for a comprehensive review of classical robust-KF methods for linear SSMs. Recently proposed provably-robust methods include the work by Boustati et al. (2020) and Cao et al. (2022).

Finally, we discuss similarities and differences with the Bayesian learning rule (BLR) line of work by Khan & Rue (2023). BLR also allows the replacing of the likelihood with a loss function, but that loss depends only on the parameter and the observation, i.e., BLR uses $\ell_t(\boldsymbol{\theta}) = L(\boldsymbol{y}_t, f_t(\boldsymbol{\theta}))$, with $L : \mathbb{R}^d \times \mathbb{R}^d$ a loss function. By contrast, our loss depends on the weight W_t, which in turn depends on the belief state, via $\hat{\boldsymbol{y}}_t = f_t(\boldsymbol{\mu}_{t|t-1})$, which in turn is a function of all the past data, i.e., our loss has the form $\ell_t(\boldsymbol{\theta}) = L(\boldsymbol{y}_{1:t}, f_t(\boldsymbol{\theta}))$. In addition, WoLF is designed for the online inference setting, whereas BLR, at least in its standard form, is designed for offline inference. Finally, note that the exponential family extension of BLR Lin et al. (2019) has also been proposed as a way to get robustness by using scale mixture posteriors. However, this induces robustness in parameter space, whereas we focus on robustness in observation space. In particular, the weighting term for the BLR mixture model depends on the distance in parameter space, $\boldsymbol{\theta}_t - \boldsymbol{\mu}_t$, whereas ours depends on the difference in observation space, $\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$.

A.1. Variational-based methods

In this section, we provide an overview of variational-Bayes (VB) robust filtering methods. As above, $\boldsymbol{\theta}_t$ is the state vector of interest and $\boldsymbol{\Psi}_t$ are additional state parameters. Given the SSM (1) and measurement model $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\Psi}_t)$, VB-based

methods seek an approximate posterior distribution over the extended state process $\mathbf{\Phi}_t = (\boldsymbol{\theta}_t, \mathbf{\Psi}_t)$ that factorises as

$$q(\mathbf{\Phi}_t) = \prod_{k=1}^{K} q(\mathbf{\Phi}_{t,k}), \tag{24}$$

with $\mathbf{\Phi}_t = (\mathbf{\Phi}_{t,1}, \ldots, \mathbf{\Phi}_{t,K})$, and K the number of collections. It can be shown that the log-density q^* that minimises the KL divergence between the true posterior distribution and the variational distribution is given by

$$\log q^*(\mathbf{\Phi}_{t,k}) = \mathbb{E}_{\neg k}\left[\log p(\boldsymbol{y}_t, \mathbf{\Phi}_t)\right] + C, \tag{25}$$

where C is the normalising constant of q^*, and the notation $\mathbb{E}_{\neg k}[\cdot]$ denotes the conditional expectation given all elements in $\mathbf{\Phi}_t$ except from $\mathbf{\Phi}_{t,k}$. See Section 10.1.1 in Bishop (2006) for details. Below, we discuss the robust VB-based filtering variants we use in the paper.

A.2. KF-IW method of Agamennoni et al. (2012)

Agamennoni et al. (2012) extend the state-space to be $\mathbf{\Psi}_t = (\boldsymbol{\theta}_t, \mathbf{R}_t)$, where $\mathbf{R}_t$ is the measurement covariance. Note that the classical KF setting, $\mathbf{R}_t$ is known. The SSM is of the form

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) &= \mathcal{N}(\boldsymbol{\theta}_t \,|\, \mathbf{F}_t\boldsymbol{\theta}_{t-1},\, \mathbf{Q}_t), \\ p(\mathbf{R}_t) &= \mathcal{W}^{-1}(\mathbf{R}_t \,|\, \nu\mathbf{\Lambda}, \nu), \\ p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \mathbf{R}_t) &= \mathcal{N}(\boldsymbol{y}_t \,|\, \mathbf{H}_t\boldsymbol{\theta}_t,\, \mathbf{R}_t), \end{aligned} \tag{26}$$

where $\mathcal{W}^{-1}(\cdot|\mathbf{P}, \eta)$ is the density of an inverse Wishart distribution with positive-definite scale matrix $\mathbf{P} \in \mathbb{R}^{m\times m}$, $\eta > m-1$ degrees of freedom, and $\nu > 0$ is the noise-scaling hyperparameter. They consider the class of variational distributions

$$q(\bar{\boldsymbol{\theta}}, \bar{\mathbf{R}}) = q(\boldsymbol{\theta}_0)q(\mathbf{R}_0)\prod_{t=1}^{T} q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})q(\mathbf{R}_t), \tag{27}$$

with $\bar{\boldsymbol{\theta}} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T)$ and $\bar{\mathbf{R}} = (\mathbf{R}_0, \ldots, \mathbf{R}_T)$. They show that the class of VB posteriors (25), under the model in (26) and (27), take the form

$$q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \mathcal{N}(\boldsymbol{\theta}_t \,|\, \boldsymbol{\mu}_t,\, \mathbf{\Sigma}_t), \tag{28}$$

$$q(\mathbf{R}_t) = \mathcal{W}^{-1}(\mathbf{R}_t|\nu_t\mathbf{\Lambda}_t, \nu_t), \tag{29}$$

with $\boldsymbol{\mu}_t$, $\mathbf{\Sigma}_t$, $\mathbf{\Lambda}_t$, and ν_t specified in Algorithm 2. In this method the hyperparameters are the number of iterations I and the

Algorithm 2 Agamennoni et al. (2012) predict and update step for i.i.d. noise with $I \geq 1$ inner iterations.

Require: $\mathbf{F}_t$, $\mathbf{Q}_t$, $\boldsymbol{\mu}_{t-1}$, $\mathbf{\Sigma}_{t-1}$ // predict step
  $\boldsymbol{\mu}_{t|t-1} \leftarrow \mathbf{F}_t\,\boldsymbol{\mu}_{t-1}$
  $\mathbf{\Sigma}_{t|t-1} \leftarrow \mathbf{F}_t\,\mathbf{\Sigma}_{t-1}\,\mathbf{F}_t^\intercal + \mathbf{Q}_t$
  $\boldsymbol{\mu}_t, \mathbf{\Sigma}_t \leftarrow \boldsymbol{\mu}_{t|t-1}, \mathbf{\Sigma}_{t|t-1}$
Require: $\boldsymbol{y}_t$, $\mathbf{H}_t$, $\mathbf{R}_t$, $\ell \in \mathbb{R}_+$ // update step
  for $i = 1, \ldots, I$ **do**
    $\mathbf{S}_t \leftarrow (\boldsymbol{y}_t - \mathbf{H}_t\,\boldsymbol{\mu}_t)(\boldsymbol{y}_t - \mathbf{H}_t\,\boldsymbol{\mu}_t)^\intercal + \mathbf{H}_t^\intercal\,\mathbf{\Sigma}_t\,\mathbf{H}_t$
    $\mathbf{\Lambda}_t \leftarrow (\ell+1)^{-1}(\ell\,\mathbf{R}_0 + \mathbf{S}_t)$
    $\mathbf{K}_t \leftarrow (\mathbf{H}_t\,\mathbf{\Sigma}_{t|t-1}\mathbf{H}_t + \mathbf{\Lambda}_t)^{-1}\mathbf{H}_t^\intercal\,\mathbf{\Sigma}_{t|t-1}$
    $\boldsymbol{\mu}_t \leftarrow \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t^\intercal(\boldsymbol{y}_t - \mathbf{H}_t^\intercal\,\boldsymbol{\mu}_{t|t-1})$
    $\mathbf{\Sigma}_t \leftarrow \mathbf{K}_t^\intercal\,\mathbf{\Lambda}_t\,\mathbf{K}_t + (\mathbf{I} - \mathbf{H}_t\,\mathbf{K}_t)^\intercal\,\mathbf{\Sigma}_{t|t-1}(\mathbf{I} - \mathbf{H}_t\,\mathbf{K}_t)$
  end for

scaling term ℓ. The prior measurement covariance is $\mathbf{R}_0$.

A.3. KF-B method of Wang et al. (2018)

Wang et al. (2018) extend the state-space to be $\boldsymbol{\Psi}_t = (\boldsymbol{\theta}_t, \xi_t, w_t)$, where w_t is an outlier event and ξ_t is its probability. The SSM is of the form:

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) &= \mathcal{N}(\boldsymbol{\theta}_t \,|\, \mathbf{F}_t\boldsymbol{\theta}_{t-1},\, \mathbf{Q}_t), \\ p(\xi_t) &= \text{Beta}(\xi_t|\alpha_0, \beta_0), \\ p(\rho_t|\xi_t) &= \text{Bern}(\rho_t|\xi_t), \\ p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \mathbf{R}_t\rho_t) &= \begin{cases} \mathcal{N}(\boldsymbol{y}_t \,|\, h_t(\boldsymbol{\theta}_t),\, \mathbf{R}_t) & \text{if } \rho_t = 1, \\ 1 & \text{if } \rho_t = 0. \end{cases} \end{aligned} \tag{30}$$

In the above equations, $\text{Beta}(\cdot|a, b)$ is the density of a Beta distribution with *shape* parameters a and b, and $\text{Bern}(\cdot|\pi)$ is the mass of a Bernoulli random variable with parameter $\pi \in [0, 1]$. They consider the class of variational distributions

$$q(\bar{\boldsymbol{\theta}}, \bar{\rho}, \bar{\xi}) = q(\boldsymbol{\theta}_0)q(\rho_0)q(\xi_0)\prod_{t=1}^{T} q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})q(\rho_t)q(\xi_t), \tag{31}$$

with $\bar{\boldsymbol{\theta}} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T)$, $\bar{\rho} = (\rho_0, \ldots, \rho_T)$, $\bar{\xi} = (\xi_0, \ldots, \xi_T)$. We provide the the predict and update equations in Algorithm 3. Here, the hyperparameters are the prior rates α_0 and β_0, and the number of inner iterations I, $\Psi(\cdot)$ is the digamma

Algorithm 3 Wang et al. (2018) predict and update step with $I \geq 1$ inner iterations.

Require: $\alpha_0, \beta_0, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}$
Require: $\mathbf{F}_t, \mathbf{Q}_t$ // predict step
  $\boldsymbol{\mu}_{t|t-1} \leftarrow \mathbf{F}_t\,\boldsymbol{\mu}_{t-1}$
  $\boldsymbol{\Sigma}_{t|t-1} \leftarrow \mathbf{F}_t\,\boldsymbol{\Sigma}_{t-1}\,\mathbf{F}_t^\intercal + \mathbf{Q}_t$
  $\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t \leftarrow \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}$
Require: $\boldsymbol{y}_t, \mathbf{H}_t, \mathbf{R}_t$, tol. $\ll 1$ // update step
  $\rho_t, \alpha', \beta' \leftarrow 1, \alpha_0, \beta_0$
  for $i = 1, \ldots, I$ **do**
    if $\rho_\epsilon <$ tol. **then**
      $\boldsymbol{\mu}_t \leftarrow \boldsymbol{\mu}_{t|t-1}$
      $\boldsymbol{\Sigma}_t \leftarrow \boldsymbol{\Sigma}_{t|-1}$
    else
      $\bar{\mathbf{R}}_t \leftarrow \mathbf{R}/\rho_t$
      $\hat{\boldsymbol{y}}_t \leftarrow \mathbf{H}_t\,\boldsymbol{\mu}_{t|t-1}$
      $\boldsymbol{\Sigma}_t^{-1} \leftarrow \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\top\,\bar{\mathbf{R}}_t^{-1}\,\mathbf{H}_t$
      $\mathbf{K}_t \leftarrow \boldsymbol{\Sigma}_t\,\mathbf{H}_t^\intercal\,\bar{\mathbf{R}}_t^{-1}$
      $\boldsymbol{\mu}_t \leftarrow \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\,(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)$
    end if
    $\mathbf{B}_t \leftarrow \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}\left[(\boldsymbol{y}_t - h_t(\boldsymbol{\theta})(\boldsymbol{y}_t - h_t(\boldsymbol{\theta}))^\intercal\right]$
    $\log\bar{\pi}_t \leftarrow \Psi(\alpha') - \Psi(\alpha' + \beta' + 1)$
    $\log(1 - \bar{\pi}_t) \leftarrow \Psi(\beta' + 1) - \Psi(\alpha' + \beta' + 1)$
    $\rho_t \leftarrow \frac{\exp(\log\bar{\pi}_t - \text{Tr}(\mathbf{B}_t\mathbf{R}_t^{-1})/2)}{\exp(\log\bar{\pi}_t - \text{Tr}(\mathbf{B}_t\mathbf{R}_t^{-1})/2) + \exp(\log(1-\bar{\pi}_t))}$
    $\alpha' \leftarrow \alpha_0 + \rho_t$
    $\beta' \leftarrow \beta_0 + 1 - \rho_t$
  end for

function, and $\mathbf{B}_t$ is of closed form after linearising the measurement function.

A.4. Ting et al. (2007)

Ting et al. (2007) extend the state-space to be $\boldsymbol{\Psi}_t = (\boldsymbol{\theta}_t, w_t)$, where w_t is a weighting term for the observation covariance $\mathbf{R}$. In their method, $\mathbf{R}$ is known and fixed. The SSM is of the form:

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) &= \mathcal{N}(\boldsymbol{\theta}_t \,|\, \mathbf{F}\boldsymbol{\theta}_{t-1},\, \mathbf{Q}), \\ p(w_t) &= \mathrm{Gam}(w_t|a_w, b_w), \\ p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) &= \mathcal{N}(\boldsymbol{y}_t \,|\, \mathbf{H}\boldsymbol{\theta}_t,\, \mathbf{R}/w_t), \end{aligned} \tag{32}$$

for a diagonal dynamics covariance $\mathbf{Q}$, $a_w, b_w > 0$, and diagonal observation covariance $\mathbf{R}$. They consider the class of variational distributions

$$q(\bar{w}, \bar{\boldsymbol{\theta}}) = q(\boldsymbol{\theta}_0) \prod_{t=1}^{T} q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) q(w_t), \tag{33}$$

with $\bar{\boldsymbol{\theta}} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T)$ and $\bar{w} = (w_0, \ldots, w_T)$. They show, for known $\mathbf{F}$, $\mathbf{H}$, $\mathbf{Q}$, and $\mathbf{R}$, that the variational distributions are of the form

$$q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \tag{34}$$

$$q(w_t) = \mathrm{Gam}(w_t|a_{w,t}, b_{w,t}), \tag{35}$$

where

$$\begin{aligned} a_{w,t} &= a_w + \frac{1}{2}, \\ b_{w,t} &= b_w + \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}[(\boldsymbol{y}_t - \mathbf{H}\boldsymbol{\theta})^\intercal \mathbf{R}^{-1}(\boldsymbol{y}_t - \mathbf{H}\boldsymbol{\theta})], \\ \boldsymbol{\Sigma}_t^{-1} &= \mathbf{Q}^{-1} + v_t \mathbf{H}^\intercal \mathbf{R}^{-1}\mathbf{H}, \\ \mathbf{K}_t &= \boldsymbol{\Sigma}_t \mathbf{H}^\intercal \mathbf{R}^{-1}, \\ \boldsymbol{\mu}_t &= \mathbf{F}\boldsymbol{\mu}_{t-1} v_t\, \mathbf{K}_t(\boldsymbol{x}_t - \mathbf{H}\mathbf{F}\boldsymbol{\mu}_{t-1}), \\ v_t &= \frac{a_{w,t} + \frac{1}{2}}{b_{w,t} + \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}[(\boldsymbol{y}_t - \mathbf{H}\boldsymbol{\theta})^\intercal \mathbf{R}^{-1}(\boldsymbol{y}_t - \mathbf{H}\boldsymbol{\theta})]}. \end{aligned} \tag{36}$$

Their method assumes no prior knowledge of either the measurement matrix $\mathbf{H}$ or the projection matrix $\mathbf{F}$. These are estimated using the EM algorithm. Assuming known $\mathbf{H}$ and $\mathbf{F}$ — as we do in this paper — allows us to bypass the M-step. However, this is detrimental to their approach since no information about the posterior covariance is propagated forward.

A.5. Huang et al. (2016)

Huang et al. (2016) extend the state-space to be $\boldsymbol{\Psi}_t = (\boldsymbol{\theta}_t, \mathbf{R}_t, \nu_t, w_t)$, where $\mathbf{R}_t$ is the measurement covariance, w_t is a weighting term for the measurement covariance, and ν_t are the degrees of freedom for the weighting term. The SSM takes the form:

$$\begin{aligned} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) &= \mathcal{N}(\boldsymbol{\theta}_t \,|\, f_t(\boldsymbol{\theta}_{t-1}),\, \mathbf{Q}_t), \\ p(\nu_t) &= \mathrm{Gam}(\nu_t \,|\, a_t, b_t), \\ p(w_t) &= \mathrm{Gam}(\lambda_t|\nu_t/2, \nu_t/2), \\ p(\mathbf{R}_t) &= \mathcal{W}^{-1}(\mathbf{R}_t \,|\, \boldsymbol{\Lambda}_t, u_t), \\ p(\boldsymbol{y}_t|\boldsymbol{\theta}_t, \mathbf{R}_t) &= \mathcal{N}(\boldsymbol{y}_t \,|\, h_t(\boldsymbol{\theta}_t),\, \mathbf{R}_t/w_t), \end{aligned} \tag{37}$$

with $\mathbf{Q}_t = \mathrm{diag}(q_{t,1}, \ldots, q_{t,D})$. Note that their method combines the SSMs in Ting et al. (2007) and Agamennoni et al. (2012).

B. WoLF-MD as a MAP estimator

In this section we show how to derive WoLF-MD as a MAP estimator. This is an alternative derivation that circumvents the use of generalised Bayes.

B.1. Deriving the Mahalanobis IMQ term

Consider the modified observation model of (32):

$$p(w_t) = \mathrm{Gam}(w_t|\alpha, \beta), \tag{38}$$

$$p(\boldsymbol{y}_t|\hat{\boldsymbol{y}}_t) = \mathcal{N}(\boldsymbol{y}_t|\hat{\boldsymbol{y}}_t, w_t^{-1}\mathbf{R}_t), \tag{39}$$

with $\hat{\boldsymbol{y}}_t = \mathbf{H}_t\boldsymbol{\mu}_{t|t-1}$ known at time t, and $\alpha, \beta > 0$. The posterior on w_t is

$$p(w_t|\boldsymbol{y}_t) \propto w_t^{\alpha-1}e^{-\beta w_t}\left|w_t\mathbf{R}_t^{-1}\right|^{1/2}\exp\left(-\frac{1}{2}\boldsymbol{e}_t^\top w_t\mathbf{R}_t^{-1}\boldsymbol{e}_t\right) \tag{40}$$

$$\propto \mathrm{Gam}\left(w_t|\alpha + \frac{n_y}{2}, \beta + \frac{1}{2}\left\|\boldsymbol{e}_t\right\|^2_{\mathbf{R}_t^{-1}}\right), \tag{41}$$

where $\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ and $\|\boldsymbol{v}\|_{\mathbf{A}} = \left\|\mathbf{A}^{1/2}\boldsymbol{v}\right\|_2$ is the Mahalanobis distance. The maximum-a-posteriori (MAP) estimate for w_t is

$$W_t = \underset{w_t\in\mathbb{R}^+}{\mathrm{argmax}}\, p(w_t|\boldsymbol{y}_t) = \frac{\alpha + \frac{n_y}{2} - 1}{\beta + \frac{1}{2}\left\|\boldsymbol{e}_t\right\|^2_{\mathbf{R}_t^{-1}}}. \tag{42}$$

For a given $c \in \mathbb{R}$, take the hyperparameters α and β to be

$$\alpha = \frac{c^2 - n_y + 2}{2}, \quad \beta = \frac{c^2}{2}, \tag{43}$$

where n_y is the number of measurements. We obtain

$$w_t = \left(1 + \frac{\left\|\boldsymbol{e}_t\right\|^2_{\mathbf{R}_t^{-1}}}{c^2}\right)^{-1/2}. \tag{44}$$

This is the Mahalanobis-based IMQ weighting function (18). Substituting the MAP estimate back into the observation model yields the weighted loglikelihood approximation

$$\log p\left(\boldsymbol{y}_t|\hat{\boldsymbol{y}}_t\right) \approx w_t^2 \log\mathcal{N}\left(\boldsymbol{y}_t|\hat{\boldsymbol{y}}_t, \mathbf{R}_t\right). \tag{45}$$

B.2. Prior Uncertainty

In this section, we take the measurement mean to be the output of a predictive model $\bar{\boldsymbol{y}}_t$ with unknown parameter $\boldsymbol{\theta}_t$. We let

$$p\left(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}\right) = \mathcal{N}\left(\boldsymbol{\theta}_t|\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right), \tag{46}$$

$$p\left(\boldsymbol{y}_t|\boldsymbol{\theta}_t, w_t\right) = \mathcal{N}\left(\boldsymbol{y}_t|\bar{\boldsymbol{y}}_t, w_t^{-1}\mathbf{R}_t\right), \tag{47}$$

where $\bar{\boldsymbol{y}}_t$ is given by (11). The joint posterior is

$$\begin{aligned}
p\left(\boldsymbol{\theta}_t, w_t|\boldsymbol{y}_{1:t}\right) \propto\; & w_t^{\alpha-1}e^{-\beta w_t}\exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}_t - \boldsymbol{\mu}_{t|t-1}\right)^\top\boldsymbol{\Sigma}_{t|t-1}^{-1}\left(\boldsymbol{\theta}_t - \boldsymbol{\mu}_{t|t-1}\right)\right) \\
& \times\left|w_t\mathbf{R}_t^{-1}\right|^{1/2}\exp\left(-\frac{1}{2}\left(\boldsymbol{y}_t - \bar{\boldsymbol{y}}_t\right)^\top w_t\mathbf{R}_t^{-1}\left(\boldsymbol{y}_t - \bar{\boldsymbol{y}}_t\right)\right) \\
\propto\; & w_t^{\alpha-1+n_y/2}\exp\left(-\frac{1}{2}\left\|\boldsymbol{\theta}_t - \boldsymbol{\mu}_{t|t-1} - \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t\mathbf{H}_t^\top\mathbf{R}_t^{-1}\mathbf{H}_t\right)^{-1}w_t\mathbf{H}_t^\top\mathbf{R}_t^{-1}\boldsymbol{e}_t\right\|^2_{\boldsymbol{\Sigma}_{t|t-1}^{-1}+w_t\mathbf{H}_t^\top\mathbf{R}_t^{-1}\mathbf{H}_t}\right) \\
& \times\exp\left(-\beta w_t - \frac{1}{2}\boldsymbol{e}_t^\top\left(w_t^{-1}\mathbf{R}_t + \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top\right)^{-1}\boldsymbol{e}_t\right).
\end{aligned}$$

where the notation $||\boldsymbol{x}||_{\mathbf{A}}$ means $\boldsymbol{x}^\intercal\,\mathbf{A}^{-1}\,\boldsymbol{x}$. Then, the marginal for w_t can be written as

$$p\left(w_t|\boldsymbol{y}_{1:t}\right) \propto w_t^{\alpha-1+n_y/2}\left|\boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t\mathbf{H}_t^\top\mathbf{R}_t^{-1}\mathbf{H}_t\right|^{-1/2}\exp\left(-\beta w_t - \frac{1}{2}\boldsymbol{e}_t^\top\left(w_t^{-1}\mathbf{R}_t + \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top\right)^{-1}\boldsymbol{e}_t\right). \tag{48}$$

By taking the limit $\boldsymbol{\Sigma}_{t|t-1} \to \mathbf{0}$, equation (48) becomes

$$\lim_{\boldsymbol{\Sigma}_{t|t-1}\to\mathbf{0}} p\left(w_t|\boldsymbol{y}_{1:t}\right) \propto w_t^{\alpha-1+n_y/2} \exp\left(-\beta w_t - \frac{1}{2}\boldsymbol{e}_t^\top \mathbf{R}_t^{-1}\boldsymbol{e}_t w_t\right), \tag{49}$$

with maximum at

$$w_t^* = \operatorname*{argmax}_{w_t} \lim_{\boldsymbol{\Sigma}_{t|t-1}\to\mathbf{0}} = \frac{\alpha - 1 + \frac{n_y}{2}}{\beta + \frac{1}{2}\boldsymbol{e}_t^\top \mathbf{R}_t^{-1}\boldsymbol{e}_t}, \tag{50}$$

that matches (42). Therefore in the SSM setting, the Mahalanobis IMQ weighting is the MAP estimate for w_t after ignoring the prior uncertainty $\boldsymbol{\Sigma}_{t|t-1}$.

C. Proofs of theoretical results

C.1. Proof of Proposition 3.1

Proof. Let $w_t^2 := W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$. The loss function takes the form

$$\begin{aligned}
\ell_t(\boldsymbol{\theta}_t) &= -w_t^2 \log \mathcal{N}(\boldsymbol{y}_t \,|\, \mathbf{H}_t\boldsymbol{\theta}_t,\, \mathbf{R}_t) \\
&= \frac{1}{2}\left(\boldsymbol{y}_t - \mathbf{H}_t\boldsymbol{\theta}_t\right)^\intercal \left(\mathbf{R}_t/w_t^2\right)^{-1}\left(\boldsymbol{y}_t - \mathbf{H}_t\boldsymbol{\theta}_t\right) - \frac{w_t^2\, d}{2}\log\pi - \frac{w_t^2}{2}\log|\mathbf{R}_t| \\
&= \frac{1}{2}\left(\boldsymbol{y}_t - \mathbf{H}_t\boldsymbol{\theta}_t\right)^\intercal \bar{\mathbf{R}}_t^{-1}\left(\boldsymbol{y}_t - \mathbf{H}_t\boldsymbol{\theta}_t\right) + C,
\end{aligned} \tag{51}$$

with $\bar{\mathbf{R}}_t = \mathbf{R}_t/w_t^2$, and where $C = -\frac{w_t^2\, d}{2}\log\pi - \frac{w_t^2}{2}\log|\mathbf{R}_t|$ is a term that does not depend on $\boldsymbol{\theta}_t$. The remaining follows from the standard KF derivation. Note that the loss function does not correspond to the log-likelihood for a homoskedastic Gaussian model since $\bar{\mathbf{R}}_t$ may depend on all data, including $\boldsymbol{y}_t$. □

Figure 8 shows the weighted log-likelihood (51) for a univariate $\mathcal{N}(0,1)$ Gaussian density as a function of the weighting term $w_t^2 \in (0,1]$. We observe that a weighting log-likelihood resembles a heavy-tailed likelihood for $w_t < 1$.

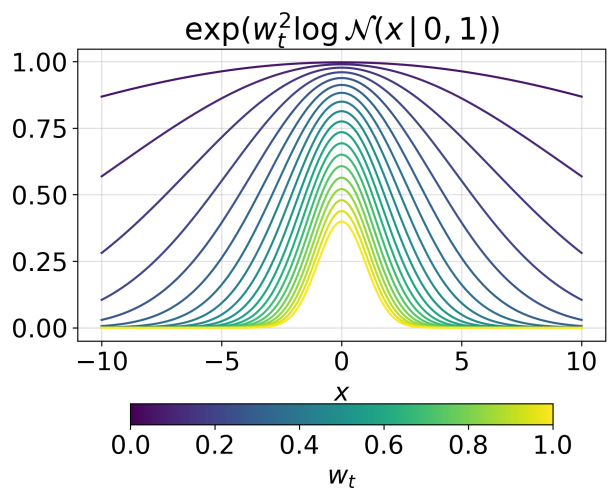

Figure 8. Weighted likelihood (unnormalised) for a standard Gaussian.

C.2. Proof of Theorem 3.2

Theorem 3.2 comprises several sub-claims, each of which we will prove as a separate lemma in this section. At the end of the section, we will integrate all the lemmas to prove the main result. The first results are for the linear Gaussian SSM case.

Lemma C.1. *Consider the linear Gaussian SSM. The standard KF posterior has an unbounded PIF and is not outlier robust.*

Proof. Let $p(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, and $p(\boldsymbol{\theta}_t|\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t^c, \boldsymbol{\Sigma}_t^c)$, the uncontaminated and contaminated standard Kalman filter posterior. Here

$$\begin{aligned} \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t\right), & \boldsymbol{\mu}_t^c &= \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t^c\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t\right), \\ \boldsymbol{\Sigma}_t^{-1} &= \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t, & (\boldsymbol{\Sigma}_t^c)^{-1} &= \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t. \end{aligned}$$

Here $\boldsymbol{\mu}_{t|t-1}, \mathbf{K}_t, \mathbf{H}_t, \boldsymbol{\Sigma}_{t|t-1}^{-1}, \mathbf{R}_t^{-1}$ do not depend on the contamination $\boldsymbol{y}_t^c$. It is clear that $\boldsymbol{\Sigma}_t^{-1} = (\boldsymbol{\Sigma}_t^c)^{-1}$, hence $\mathbf{K}_t^c = \boldsymbol{\Sigma}_t^c\mathbf{H}_t^\intercal\mathbf{R}_t^{-1} = \boldsymbol{\Sigma}_t\mathbf{H}_t^\intercal\mathbf{R}_t^{-1} = \mathbf{K}_t$. Using the fact that given two d-dimensional Gaussians $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, the KL divergence is

$$\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\|\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) = \frac{1}{2}\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) - m + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \ln\left(\frac{\det \boldsymbol{\Sigma}_1}{\det \boldsymbol{\Sigma}_0}\right)\right),$$

and given that $\mathrm{Tr}(\mathbf{I}_p) = p$, we can derive the PIF as

$$\begin{aligned} \mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}) &= \mathrm{KL}\left(p(\boldsymbol{\theta}_t|\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1})\|p(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1})\right) \\ &= \frac{1}{2}\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_t\right) - m + (\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c)^\intercal \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c) + \ln\left(\frac{\det \boldsymbol{\Sigma}_t}{\det \boldsymbol{\Sigma}_t}\right)\right) \\ &= \mathrm{Tr}\left(\mathbf{I}_m\right) - m + \frac{1}{2}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c)^\intercal \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c) \\ &= \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{y}_t^c)^\intercal \mathbf{K}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\mathbf{K}_t(\boldsymbol{y}_t - \boldsymbol{y}_t^c). \end{aligned}$$

Here, $\mathbf{K}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\mathbf{K}_t$ is positive definite. To see this, first note that for every invertible matrix B and conformable positive definite matrix A, $B^\intercal AB$ is positive definite (Chapter 7.1 Horn & Johnson, 2012). Since $\mathbf{K}_t$ is a product of invertible matrices, it therefore is itself invertible. In fact, let $\boldsymbol{z} \in \mathbb{R}^m$ be a non-zero vector. We can write $\boldsymbol{z}^\intercal B^\intercal AB\boldsymbol{z} = \tilde{\boldsymbol{z}}^\intercal A\tilde{\boldsymbol{z}}$, for $\tilde{\boldsymbol{z}} = B\boldsymbol{z}$. We know that $\tilde{\boldsymbol{z}} = B\boldsymbol{z} \neq 0$, because B is invertible. Then, since A is positive definite and $\tilde{\boldsymbol{z}} \in \mathbb{R}^m$, it holds that $\tilde{\boldsymbol{z}}^\intercal A\tilde{\boldsymbol{z}} > 0$, proving that $B^\intercal AB$ is positive definite. Finally, since $\boldsymbol{\Sigma}_t^{-1}$ is positive definite, it therefore holds that $\mathbf{K}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\mathbf{K}_t$ is positive definite, too.

This in turn allows us to lower-bound the PIF by

$$\frac{1}{2}\lambda_{\min}(\mathbf{K}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\mathbf{K}_t)\|\boldsymbol{y}_t - \boldsymbol{y}_t^c\|_2^2 \leq \frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{y}_t^c)^\intercal \mathbf{K}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\mathbf{K}_t(\boldsymbol{y}_t - \boldsymbol{y}_t^c) = \mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}),$$

where we use the fact that for every positive definite matrix A and vector $\boldsymbol{z}$, we have $\lambda_{\min}(A)\|\boldsymbol{z}\|^2 \leq \boldsymbol{z}^\intercal A\boldsymbol{z} \leq \lambda_{\max}(A)\|\boldsymbol{z}\|^2$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalues of A, respectively (Chapter 5.7 Horn & Johnson, 2012). Moreover, we know that $\lambda_{\min}\left(\mathbf{K}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\mathbf{K}_t\right) > 0$, since $\mathbf{K}_t^\intercal\boldsymbol{\Sigma}_t^{-1}\mathbf{K}_t$ is positive definite, and it does not depend on $\boldsymbol{y}_t^c$. Therefore, it indeed holds that $\mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}) \to +\infty$ as $\|\boldsymbol{y}_t - \boldsymbol{y}_t^c\|_2^2 \to +\infty$. □

Lemma C.2. *Consider the linear Gaussian SSM. The generalised posterior presented in Proposition 3.1 has bounded PIF and is, therefore, outlier robust for any weighting function W such that $\sup_{\boldsymbol{y}_t \in \mathbf{R}^d} W(\boldsymbol{y}_{1:t}) < \infty$ and $\sup_{\boldsymbol{y}_t \in \mathbf{R}^d} W(\boldsymbol{y}_{1:t})^2\|\boldsymbol{y}_t\|_2 < \infty$.*

Proof. Let W be a weighting function such that $\sup_{\boldsymbol{y}_t \in \mathbb{R}^d} W(\boldsymbol{y}_{1:t}) < \infty$ and $\sup_{\boldsymbol{y}_t \in \mathbb{R}^d} W(\boldsymbol{y}_{1:t})^2\|\boldsymbol{y}_t\|_2 < \infty$. Lets define $w_t := W^2(\boldsymbol{y}_{1:t})$, and $w_t^c := W^2(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:1-t})$. Let $q(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, and $q(\boldsymbol{\theta}_t|\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t^c, \boldsymbol{\Sigma}_t^c)$, where

$$\begin{aligned} \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t|t-1} + w_t\mathbf{K}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t\right), & \boldsymbol{\mu}_t^c &= \boldsymbol{\mu}_{t|t-1} + w_t^c\mathbf{K}_t^c\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t\right), \\ \boldsymbol{\Sigma}_t^{-1} &= \boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t\mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t, & (\boldsymbol{\Sigma}_t^c)^{-1} &= \boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t^c\mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t, \end{aligned}$$

as in Algorithm 1 Then, the PIF has the form

$$\mathrm{PIF}(\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1}) = \frac{1}{2}\left(\underbrace{\mathrm{Tr}\left((\boldsymbol{\Sigma}_t^c)^{-1}\boldsymbol{\Sigma}_t\right) - p}_{(1)} + \underbrace{(\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c)^\intercal (\boldsymbol{\Sigma}_t^c)^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c)}_{(2)} + \underbrace{\ln\left(\frac{\det \boldsymbol{\Sigma}_t^c}{\det \boldsymbol{\Sigma}_t}\right)}_{(3)}\right).$$

Now, we will get a bound for each term in the PIF. The first term can be bounded as

$$(1) = \operatorname{Tr}\left((\boldsymbol{\Sigma}_t^c)^{-1}\boldsymbol{\Sigma}_t\right) - m \leq \operatorname{Tr}\left((\boldsymbol{\Sigma}_t^c)^{-1}\right)\operatorname{Tr}\left(\boldsymbol{\Sigma}_t\right) - m,$$

where we use the fact that for two positive semidefinite matrices A, B, it holds that $\operatorname{Tr}(AB) \leq \operatorname{Tr}(A)\operatorname{Tr}(B)$. Observing that $\boldsymbol{\Sigma}_t$ does not depend on y_t^c, we can now write $C_1 = \boldsymbol{\Sigma}_t$, so that by using the arithmetic rules of traces, we obtain

$$\begin{aligned}(1) &\leq C_1 \operatorname{Tr}\left((\boldsymbol{\Sigma}_t^c)^{-1}\right) - m \\ &= C_1 \operatorname{Tr}\left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t^c \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t\right) - m \\ &= C_1 \operatorname{Tr}\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\right) + C_1 w_t^c \operatorname{Tr}\left(\mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t\right) - m.\end{aligned}$$

Here, we use the fact that the trace of a sum is a sum of traces, and the trace of the constant times matrix is equal to the constant times trace of the matrix. Finally, since $\sup_{\boldsymbol{y}_t^c \in \mathbb{R}^d} w_t^c \leq C_2 < \infty$, the entire expression can be bounded by a constant $C_3 < \infty$ that does not depend on the contamination $\boldsymbol{y}_t^c$ as

$$(1) \leq C_1 \operatorname{Tr}\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\right) + C_1 C_2 \operatorname{Tr}\left(\mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t\right) - m = C_3,$$

where we use the fact that both traces are finite since both matrices are real-valued. Next, we bound the second term by noting that it is the squared Mahalanobis norm of $\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c$ with respect to $\boldsymbol{\Sigma}_t^c$, and we write it as $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c\|_{\boldsymbol{\Sigma}_t^c}$ (in particular, it then satisfies all the properties of a norm). Therefore, we apply the triangle inequality to obtain

$$\begin{aligned}(2) &= (\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c\|_{\boldsymbol{\Sigma}_t^c})^2 \\ &= (\|w_t \mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t) - w_t^c \mathbf{K}_t^c(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c})^2 \\ &\leq (\|w_t \mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c} + \|w_t^c \mathbf{K}_t^c(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c})^2 \\ &\leq 2(\|w_t \mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c}^2 + \|w_t^c \mathbf{K}_t^c(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c}^2).\end{aligned}$$

In the last inequality, we use the fact that, for two real numbers a and b, it holds that $(a+b)^2 \leq 2(a^2+b^2)$. Then, we bound each term separately,

$$\begin{aligned}2\|w_t \mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c}^2 &= 2w_t^2(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)^\intercal \mathbf{K}_t^\intercal (\boldsymbol{\Sigma}_t^c)^{-1}\mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t) \\ &\leq 2w_t^2 \lambda_{\max}\left(\mathbf{K}_t^\intercal (\boldsymbol{\Sigma}_t^c)^{-1}\mathbf{K}_t\right)\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t\|_2^2,\end{aligned}$$

where $2\lambda_{\max}\left(\mathbf{K}_t^\intercal (\boldsymbol{\Sigma}_t^c)^{-1}\mathbf{K}_t\right) \leq C_4 < \infty$, since $\mathbf{K}_t^\intercal (\boldsymbol{\Sigma}_t^c)^{-1}\mathbf{K}_t$ is a positive definite matrix, and thus, all eigenvalues are real-valued. This property arises from the fact that for every invertible matrix B and a positive definite matrix A, $B^\intercal AB$ is positive definite. Finally, we must verify that $(\boldsymbol{\Sigma}_t^c)^{-1}$ is positive definite. Let $\boldsymbol{z} \in \mathbb{R}^m$ be a non-zero vector. Then,

$$\boldsymbol{z}^\intercal\left(\boldsymbol{\Sigma}_t^c\right)^{-1}\boldsymbol{z} = \boldsymbol{z}^\intercal\left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t^c \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t\right)\boldsymbol{z} = \boldsymbol{z}^\intercal\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\right)\boldsymbol{z} + w_t^c\left(\boldsymbol{z}^\intercal \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t\boldsymbol{z}\right).$$

We know that $\boldsymbol{\Sigma}_{t|t-1}^{-1}$ is positive definite, hence $\boldsymbol{z}^\intercal\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\right)\boldsymbol{z} > 0$. Moreover, $\boldsymbol{z}^\intercal \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t\boldsymbol{z} > 0$ because $\mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t$ is positive definite. Since $w_t^c = W^2(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:1-t}) \geq 0$ for all $\boldsymbol{y}_t^c \in \mathbb{R}^d$, then $w_t^c\left(\boldsymbol{z}^\intercal \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t\boldsymbol{z}\right) \geq 0$. Finally, combining these inequalities, $\boldsymbol{z}^\intercal\left(\boldsymbol{\Sigma}_t^c\right)^{-1}\boldsymbol{z} > 0$. Therefore, $(\boldsymbol{\Sigma}_t^c)^{-1}$ is positive definite. Therefore,

$$2\|w_t \mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c}^2 \leq C_4(w_t\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t\|_2)^2 = C_5$$

since $(w_t\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t\|_2)^2$ does not depend on the contamination. Similarly,

$$2\|w_t^c \mathbf{K}_t^c(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t)\|_{\boldsymbol{\Sigma}_t^c}^2 \leq C_6(w_t^c\|\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t\|_2)^2,$$

where, using the same argument as before, $2\lambda_{\max}\left((\mathbf{K}_t^c)^\intercal (\boldsymbol{\Sigma}_t^c)^{-1}\mathbf{K}_t^c\right) \leq C_6 < \infty$. Now, since $\sup_{\boldsymbol{y}_t^c \in \mathbf{R}^d} w_t^c\|\boldsymbol{y}_t^c\|_2 \leq C_7 < \infty$, we have:

$$C_6(w_t^c\|\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t\|_2)^2 \leq C_6(w_t^c\|\boldsymbol{y}_t^c\|_2 + w_t^c\|\hat{\boldsymbol{y}}_t\|_2)^2 \leq C_6(C_7 + C_2\|\hat{\boldsymbol{y}}_t\|_2)^2 = C_8.$$

Putting it all together, we find that

$$(2) \leq C_5 + C_8.$$

Lastly, the third and final term can be rewritten using properties of determinants as

$$(3) = \ln\left(\frac{\det \boldsymbol{\Sigma}_t^c}{\det \boldsymbol{\Sigma}_t}\right) = \ln\left(\frac{1}{\det \boldsymbol{\Sigma}_t}\right) + \ln\left(\det \boldsymbol{\Sigma}_t^c\right) = \ln\left(\frac{1}{\det \boldsymbol{\Sigma}_t}\right) + \ln\left(\frac{1}{\det(\boldsymbol{\Sigma}_t^c)^{-1}}\right).$$

We define $C_9 = \ln\left(\frac{1}{\det \boldsymbol{\Sigma}_t}\right)$ since it does not depend on the contamination, and write

$$\begin{aligned}
(3) &= C_9 + \ln\left(\frac{1}{\det(\boldsymbol{\Sigma}_t^c)^{-1}}\right) \\
&= C_9 + \ln\left(\frac{1}{\det\left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t^c \mathbf{H}_t^\intercal \mathbf{R}_t^{-1} \mathbf{H}_t\right)}\right) \\
&\leq C_9 + \ln\left(\frac{1}{\det\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\right) + \det\left(w_t^c \mathbf{H}_t^\intercal \mathbf{R}_t^{-1} \mathbf{H}_t\right)}\right),
\end{aligned}$$

where in the last inequality, we use the fact that for two positive semidefinite matrices A, B, it also holds that $\det(A+B) \geq \det(A) + \det(B)$. Finally,

$$\begin{aligned}
(3) &\leq C_9 + \ln\left(\frac{1}{\det\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\right) + \det\left(w_t^c \mathbf{H}_t^\intercal \mathbf{R}_t^{-1} \mathbf{H}_t\right)}\right) \\
&\leq C_9 + \ln\left(\frac{1}{\det\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\right)}\right) = C_{10}.
\end{aligned}$$

Here, we use the fact that $w_t^c \mathbf{H}_t^\intercal \mathbf{R}_t^{-1} \mathbf{H}_t$ is positive semidefinite, as we showed previously. Therefore $\det\left(w_t^c \mathbf{H}_t^\intercal \mathbf{R}_t^{-1} \mathbf{H}_t\right) \geq 0$. By putting the bounds for (1), (2), and (3) together, we obtain

$$\begin{aligned}
\mathrm{PIF}(\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1}) &= \frac{1}{2}\Bigg(\underbrace{\mathrm{Tr}\left((\boldsymbol{\Sigma}_t^c)^{-1}\boldsymbol{\Sigma}_t\right) - d}_{(1)} + \underbrace{(\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c)^\intercal (\boldsymbol{\Sigma}_t^c)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{\mu}_t^c)}_{(2)} + \underbrace{\ln\left(\frac{\det \boldsymbol{\Sigma}_t^c}{\det \boldsymbol{\Sigma}_t}\right)}_{(3)}\Bigg). \\
&\leq C_3 + C_5 + C_8 + C_{10} < \infty.
\end{aligned}$$

□

We now extend the result to the linearised approximation of the SSM case.

Lemma C.3. *Consider the linearised approximation of the SSM. The standard EKF posterior has an unbounded PIF and is not outlier robust.*

Proof. We can easily replicate the procedure in Lemma C.1 to the EKF since it can be straightforwardly applied to the approximate posterior presented in Section 2.2. Specifically, for the standard EKF, we compute:

$$\mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}) = \mathrm{KL}\left(p(\boldsymbol{\theta}_t|\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1}) \| p(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1})\right) \tag{52}$$

for $p(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, and $p(\boldsymbol{\theta}_t|\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t^c, \boldsymbol{\Sigma}_t^c)$, representing the uncontaminated and contaminated standard EKF posterior. In particular,

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t), \qquad \boldsymbol{\mu}_t^c = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t^c(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t),$$
$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t, \qquad (\boldsymbol{\Sigma}_t^c)^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t.$$

Here, $\boldsymbol{\mu}_{t|t-1}, \mathbf{K}_t, \mathbf{H}_t, \boldsymbol{\Sigma}_{t|t-1}^{-1}, \mathbf{R}_t^{-1}$ do not depend on the contamination $\boldsymbol{y}_t^c$, and $\boldsymbol{\mu}_{t|t-1} = \mathbb{E}[\bar{\boldsymbol{\mu}}_{t|t-1}|\boldsymbol{\mu}_{t-1}] = f_t(\boldsymbol{\mu}_{t-1})$, $\hat{\boldsymbol{y}}_t = \mathbb{E}[\bar{\boldsymbol{y}}_t] = h_t(\boldsymbol{\mu}_{t|t-1})$, where $\bar{\boldsymbol{\mu}}_{t|t-1}$ and $\bar{\boldsymbol{y}}_t$ are defined as Equation (10) and Equation (11) respectively, and $\mathbf{H}_t$ is the Jacobian of h_t evaluated at $\boldsymbol{\mu}_{t|t-1}$. Since it follows the same structure as the standard Kalman filter, replicating the procedure in Lemma C.1, we obtain that the standard EKF is not robust. □

Lemma C.4. *Consider the linearised approximation of the SSM. The generalised posterior presented in Proposition 3.1 has bounded PIF and is, therefore, outlier robust for any weighting function W such that* $\sup_{\boldsymbol{y}_t \in \mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) < \infty$ *and* $\sup_{\boldsymbol{y}_t \in \mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)^2 \|\boldsymbol{y}_t\|_2 < \infty$.

Proof. We can easily replicate the procedure in Lemma C.2 since it can be straightforwardly applied to the approximate posterior presented in Section 3.2. Now consider the weighted EKF. We compute

$$\mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}) = \mathrm{KL}\left(q(\boldsymbol{\theta}_t|\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1}) \| q(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1})\right). \tag{53}$$

Let $W : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a weighting function such that $\sup_{\boldsymbol{y}_t \in \mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) < \infty$ and $\sup_{\boldsymbol{y}_t \in \mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)^2 \|\boldsymbol{y}_t\|_2 < \infty$. Define $w_t := W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$, and $w_t^c := W^2(\boldsymbol{y}_t^c, \hat{\boldsymbol{y}}_t)$.

Let $q(\boldsymbol{\theta}_t|\boldsymbol{y}_t, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, and $q(\boldsymbol{\theta}_t|\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t-1}) = \mathcal{N}(\boldsymbol{\theta}_t|\boldsymbol{\mu}_t^c, \boldsymbol{\Sigma}_t^c)$, where

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + w_t\mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t), \qquad \boldsymbol{\mu}_t^c = \boldsymbol{\mu}_{t|t-1} + w_t^c\mathbf{K}_t^c(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_t),$$
$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t\mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t, \qquad (\boldsymbol{\Sigma}_t^c)^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} + w_t^c\mathbf{H}_t^\intercal \mathbf{R}_t^{-1}\mathbf{H}_t.$$

for $\boldsymbol{\mu}_{t|t-1}$ and $\hat{\boldsymbol{y}}_t$ as before. Then, following the same procedure as in the proof of Lemma C.2, we show that this PIF is bounded and the method is robust. □

Proof of Theorem 3.2 The proof of Theorem 3.2 follows directly from Lemmas C.1 to C.4.

C.3. Ensemble Kalman Filter

Now, we extend Theorem 3.2 to the ensemble Kalman filter case. There are a couple of caveats to proving robustness for this case. While we know the distribution of the particles in the case where the state-space model is linear and Gaussian, this is not always the case. Therefore, we propose studying the empirical measures defined by the particles. The problem with this is that the Kullback-Leibler divergence is not defined for empirical measures with different supports. This is why we propose using the 2-Wasserstein distance instead since it is unbounded and well-defined for empirical measures. The 2-Wasserstein distance is defined as follows: If P is an empirical measure with samples $x_1, ..., x_n$ and Q is an empirical measure with samples $y_1, ..., y_n$,

$$D_{W_2}(P, Q) = \inf_{\pi} \left(\frac{1}{n} \sum_{i=1}^{n} \left\| x_i - y_{\pi(i)} \right\|_2^2 \right)^{1/2},$$

where the infimum is over all permutations π of n elements. Therefore, the PIF for the ensemble Kalman filter case has the form

$$\mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}) = D_{W_2}\left(\mathbb{P}_N, \mathbb{P}_N^c\right), \tag{54}$$

where $\mathbb{P}_N$ is the empirical measure of the (non-contaminated) particles $\left\{\hat{\theta}_t^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + w_t\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right)\right\}_{i=1}^N$, and $\mathbb{P}_N^c$ is the empirical measure of the contaminated particles $\left\{\hat{\theta}_{t^c}^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right)\right\}_{i=1}^N$. As in the previous definition, if $\sup_{\boldsymbol{y}_t^c \in \mathbb{R}^d} |\mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t})| < \infty$, then the posterior is called outlier-robust.

Lemma C.5. *The generalised posterior presented in Appendix D.3 has bounded PIF (as defined in Equation (54)) and is, therefore, outlier robust for any weighting function* $W : \mathbb{R}^{d\times d} \to \mathbb{R}$ *such that* $\sup_{\boldsymbol{y}_t\in\mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) < \infty$ *and* $\sup_{\boldsymbol{y}_t\in\mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)^2 \|\boldsymbol{y}_t\|_2 < \infty$.

Proof. Let $W : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a weighting function such that $\sup_{\boldsymbol{y}_t\in\mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) < \infty$ and $\sup_{\boldsymbol{y}_t\in\mathbf{R}^d} W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)^2\|\boldsymbol{y}_t\|_2 < \infty$. Define $w_t := W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$, and $w_t^c := W^2(\boldsymbol{y}_t^c, \hat{\boldsymbol{y}}_t)$. Then, the PIF

$$\mathrm{PIF}(\boldsymbol{y}_t^c, \boldsymbol{y}_{1:t}) = D_{W_2}(\mathbb{P}_N, \mathbb{P}_N^c) = \inf_{\pi} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\theta}_t^{(i)} - \hat{\theta}_{t^c}^{(\pi(i))} \right\|_2^2 \right)^{1/2},$$

where the infimum is over all permutations π of N elements, $\mathbb{P}_N$ is the empirical measure of the particles $\left\{\hat{\theta}_t^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + w_t\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right)\right\}_{i=1}^N$, and $\mathbb{P}_N^c$ is the empirical measure of the contaminated particles $\left\{\hat{\theta}_{t^c}^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right)\right\}_{i=1}^N$. Therefore, we know that the infimum will be smaller than considering only the identity as permutation,

$$D_{W_2}(\mathbb{P}_N, \mathbb{P}_N^c) \leq \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\theta}_t^{(i)} - \hat{\theta}_{t^c}^{(i)} \right\|_2^2 \right)^{1/2}.$$

Now, using the equation of the particles:

$$\begin{aligned} D_{W_2}(\mathbb{P}_N, \mathbb{P}_N^c) &\leq \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + w_t\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) - \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} - w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 \right)^{1/2} \\ &= \left(\frac{1}{N} \sum_{i=1}^{N} \left\| w_t\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) - w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 \right)^{1/2}. \end{aligned}$$

We use triangle inequality to obtain:

$$\begin{aligned} D_{W_2}(\mathbb{P}_N, \mathbb{P}_N^c) &\leq \left(\frac{1}{N} \sum_{i=1}^{N} \left(\left\| w_t\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2 + \left\| w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2 \right)^2 \right)^{1/2} \\ &\leq \left(\frac{2}{N} \sum_{i=1}^{N} \left\| w_t\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 + \left\| w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 \right)^{1/2}. \end{aligned}$$

Lets define $C_{11} = \sum_{i=1}^N \|w_t\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right)\|_2^2$, a constant that does not depend on the contamination. Therefore:

$$D_{W_2}(\mathbb{P}_N, \mathbb{P}_N^c) \leq \left(\frac{2}{N} \left(C_{11} + \sum_{i=1}^{N} \left\| w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 \right) \right)^{1/2}.$$

First, we observe that $\left\| w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 = (w_t^c)^2 \left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right)^\intercal \bar{\mathbf{K}}_t^\intercal\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right)$. Now we bound this expression using the fact that for every positive definite matrix A and vector $\boldsymbol{z}$, we have $\lambda_{\min}(A)\|\boldsymbol{z}\|^2 \leq \boldsymbol{z}^\intercal A\boldsymbol{z} \leq \lambda_{\max}(A)\|\boldsymbol{z}\|^2$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalues of A, respectively (Chapter 5.7 Horn & Johnson, 2012):

$$\left\| w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 \leq (w_t^c)^2\lambda_{\max}(\bar{\mathbf{K}}_t^\intercal\bar{\mathbf{K}}_t) \left\| \boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)} \right\|_2^2,$$

where $\lambda_{\max}\left(\bar{\mathbf{K}}_t^\intercal\bar{\mathbf{K}}_t\right) \leq C_{12} < \infty$, since $\bar{\mathbf{K}}_t^\intercal\bar{\mathbf{K}}_t$ is a positive definite matrix, and thus, all eigenvalues are real-valued. Then,

$$\left\| w_t^c\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right) \right\|_2^2 \leq C_{12}\left(w_t^c \left\| \boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)} \right\|_2 \right)^2.$$

Now, since $\sup_{\boldsymbol{y}_t^c \in \mathbb{R}^d} w_t^c \leq C_{13} < \infty$, and $\sup_{\boldsymbol{y}_t^c \in \mathbf{R}^d} w_t^c \|\boldsymbol{y}_t^c\|_2 \leq C_{14} < \infty$, we have:

$$C_{12}\left(w_t^c \left\|\boldsymbol{y}_t^c - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right\|_2\right)^2 \leq C_{12}\left(w_t^c \|\boldsymbol{y}_t^c\|_2 + w_t^c \left\|\hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right\|_2\right)^2 \leq C_{12}\left(C_{14} + C_{13}\left\|\hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right\|_2\right)^2 = C_{15}.$$

Putting it all together, we find that

$$D_{W_2}(\mathbb{P}_N, \mathbb{P}_N^c) \leq \left(\frac{2}{N}\left(C_{11} + \sum_{i=1}^{N} C_{15}\right)\right)^{1/2} < \infty.$$

□

D. Extensions of the weighted likelihood filter

Below, we discuss a number of extension to the WoLF methodology including generalisations to (i) exponential-family members, (ii) multi-output weighting functions, and (iii) the EnKF.

D.1. Exponential family likelihoods

We extend WoLF for measurements modelled using an element of the exponential family of distributions (as first mentioned in Section 3.2). Classical examples of exponential families, in addition to the Gaussian distribution, are the Bernoulli distribution, the Gamma distribution, the Beta distribution, and the Poisson distribution. These distributions can be considered to tackle filtering problems when the measurements are generated respectively from a binary process, a process that only takes values in positive real line, a process that takes values in the interval $[0, 1]$, or a counting process.

We take the mass function of a measurement $\boldsymbol{y}_t \in B \subseteq \mathbb{R}^d$ to be of the form

$$p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \text{expfam}(\boldsymbol{y}_t|\boldsymbol{\eta}_t) = Z^{-1}(\boldsymbol{\eta}_t)\exp\left(\boldsymbol{\eta}_t^\intercal T(\boldsymbol{y}_t) + b(\boldsymbol{y}_t)\right). \tag{55}$$

with B the support of the measurement $\boldsymbol{y}_t$, $\boldsymbol{\eta}_t \in \mathbb{R}^k$ the natural parameters, $A : B \to \mathbb{R}^k$ the sufficient statistic function, $b : B \to \mathbb{R}$ the base measure, and $Z : \mathbb{R}^k \to \mathbb{R}$ the normalising function. To simplify the notation, let $\boldsymbol{\eta}_t = h_t(\boldsymbol{\theta}_t)$ and $\boldsymbol{z}_t = T(\boldsymbol{y}_t)$. We also define the dual (moment) parameters

$$\boldsymbol{\lambda}_t = \mathbb{E}\left[\boldsymbol{z}_t|\boldsymbol{\eta}_t\right] = \nabla_{\boldsymbol{\eta}_t} \log Z(\boldsymbol{\eta}_t) \tag{56}$$

and the conditional variance

$$\mathbf{R}_t = \text{Cov}[\boldsymbol{z}_t|\boldsymbol{\eta}_t] = \nabla_{\boldsymbol{\eta}_t}\boldsymbol{\lambda}_t = \nabla^2_{\boldsymbol{\eta}_t} \log Z(\boldsymbol{\eta}_t), \tag{57}$$

and we take our predictive model to output the dual parameters:

$$\boldsymbol{\lambda}_t = h_t(\boldsymbol{\theta}_t). \tag{58}$$

For example, for a Gaussian likelihood with fixed observation noise and a linear observation model as in (5), we have that $A(\boldsymbol{y}_t) = \boldsymbol{y}_t$, $\boldsymbol{\lambda}_t = \mathbf{H}_t\boldsymbol{\theta}_t$, and $\mathbf{R}_t$ is constant.

The exponential family EKF algorithm of Ollivier (2018) approximates the likelihood in (55) with a moment-matched Gaussian,

$$q(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \mathcal{N}(\boldsymbol{z}_t|\bar{\boldsymbol{\lambda}}_t, \mathbf{R}_t), \tag{59}$$

where $\bar{\boldsymbol{\lambda}}_t = \mathbf{H}_t(\boldsymbol{\theta}_t - \bar{\boldsymbol{\mu}}_{t|t-1}) + h_t(\bar{\boldsymbol{\mu}}_{t|t-1})$, and $\boldsymbol{\mu}_{t|t-1}$ is given by (10). We modify this algorithm by including a weighting term on the log-likelihood to get

$$\log q(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = W_t(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)^2\left(-\frac{1}{2}\bar{\boldsymbol{\lambda}}_t^\top \mathbf{R}_t^{-1}\bar{\boldsymbol{\lambda}}_t + \bar{\boldsymbol{\lambda}}_t^\top \mathbf{R}_t^{-1} T(\boldsymbol{y}_t)\right) + \text{cst}, \tag{60}$$

with $\text{cst}.$ a constant term that does not depend on $\boldsymbol{\theta}$. We leave the study of $W(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$ when modelling a non-Gaussian exponential family for future work.

D.2. Dimension-specific weighting

In this section, we provide details about the implementation of a vector-valued weighting function $W_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ introduced in Section 3.3. See Section 4.3 for an evaluation of this method. We consider $\boldsymbol{y}_t \in \mathbb{R}^d$ with $d > 1$.

If the likelihood factorises as $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \prod_{j=1}^d p(y_{t,j}|\boldsymbol{\theta}_t)$, with $y_{t,j}$ the j-th element of the t-th measurement, we define the weighted log-likelihood as

$$\log q(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \sum_{j=1}^{d} w_{t,j}^2 \log p(y_{t,j}|\boldsymbol{\theta}_t), \tag{61}$$

where $w_{t,j}^2 = W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)_j$ is the j-th entry of the vector-valued weight function.

For a Gaussian likelihood $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \mathcal{N}(\boldsymbol{y}_t|h_t(\boldsymbol{\theta}_t), \mathbf{R}_t)$ we define the weighted likelihood as

$$q(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \mathcal{N}\left(\boldsymbol{y}_t|h_t(\boldsymbol{\theta}_t), \bar{\mathbf{R}}_t\right), \tag{62}$$

$$\bar{\mathbf{R}}_t^{-1} = \mathrm{Diag}(\boldsymbol{w}_t)\mathbf{R}_t^{-1}\mathrm{Diag}(\boldsymbol{w}_t), \tag{63}$$

which is a special case of (61) when $\mathbf{R}_t$ is diagonal.[8] This expression scales the precision of each $y_{t,j}$ by $w_{t,j}^2$ while preserving correlations. It can also be written in terms of dimension-specific weighting on the errors:

$$\log q(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = -\frac{1}{2}\tilde{E}_t(\boldsymbol{\theta}_t)^\intercal \mathbf{R}_t^{-1}\tilde{E}_t(\boldsymbol{\theta}_t), \tag{64}$$

$$\tilde{E}_t(\boldsymbol{\theta}_t) = \mathrm{Diag}(\boldsymbol{w}_t)(\boldsymbol{y}_t - h_t(\boldsymbol{\theta}_t)). \tag{65}$$

The EKF update from the weighted loglikelihood in (62) mirrors the standard EKF update with the true observation precision $\mathbf{R}_t^{-1}$ replaced by $\bar{\mathbf{R}}_t^{-1}$:

$$\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^\top \bar{\mathbf{R}}_t^{-1}\mathbf{H}_t, \tag{66}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t), \tag{67}$$

$$\mathbf{K}_t = \boldsymbol{\Sigma}_t\mathbf{H}_t^\top \bar{\mathbf{R}}_t^{-1}. \tag{68}$$

This generalises the WoLF update with scalar weight given by Proposition 3.1.

Finally, when $w_{t,j} = 0$ for one or more observation dimensions, (61) and (62) define improper observation distributions because the marginals on these dimensions are uniform over $\mathbb{R}$ and the precision $\bar{\mathbf{R}}_t^{-1}$ is singular. Nevertheless, the joint marginal for the dimensions with positive weights, $\boldsymbol{y}_{t,J_t}$ where $J_t = \{j : w_{t,j} > 0\}$, is a proper distribution. This is all that is needed because the observations $y_{t,j}$ with $w_{t,j} = 0$ are ignored in the update. This can be seen in (66), (67), and (68), where as a consequence of (63), the error terms and the Jacobian are both zeroed out on the zero-weighted dimensions.

D.3. Weighted ensembles

In this section, we introduce the weighted EnKF first mentioned in Section 3.2. We begin with a discussion of the EnKF originally referenced in Section 2.3.

The ensemble Kalman filter: The ensemble Kalman filter (EnKF) (Roth et al., 2017b) uses an ensemble of $N \in \mathbb{N}$ particles $\{\hat{\boldsymbol{\theta}}_{t|t-1}^{(i)}\}_{i=1}^N$. For each $i = 1, \ldots, N$, the update step samples predictions $\hat{\boldsymbol{y}}_{t|t-1}^{(i)}$ according to

$$\hat{\boldsymbol{y}}_{t|t-1}^{(i)} \sim \mathcal{N}\left(h_t\left(\hat{\boldsymbol{\theta}}_{t|t-1}^{(i)}\right), \mathbf{R}_t\right). \tag{69}$$

and then updates each particle according to

$$\hat{\boldsymbol{\theta}}_t^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + \bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right), \tag{70}$$

The EnKF gain $\bar{\mathbf{K}}_t$ is

$$\bar{\mathbf{K}}_t = \mathrm{cov}_i\left[\hat{\boldsymbol{\theta}}_{t|t-1}^{(i)}, \hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right]\left(\mathbb{V}_i\left[\hat{\boldsymbol{y}}_{t|t-1}^{(i)}\right]\right)^{-1}. \tag{71}$$

[8] $\mathrm{Diag}(\boldsymbol{v})$ for $\boldsymbol{v} \in \mathbb{R}^K$ is the $K \times K$ matrix with $\mathrm{Diag}(\boldsymbol{v})_{i,i} = \boldsymbol{v}_i$ and $\mathrm{Diag}(\boldsymbol{v})_{i,j} = 0$ for $i \neq j$.

We write $\mathbb{E}_i[\cdot], \mathrm{cov}_i[\cdot,\cdot], \mathbb{V}_i[\cdot]$ to refer to the distribution over particles as indexed by i.

The EnKF converges to standard KF as $N \to \infty$ when the likelihood is linear-Gaussian and the prior is also Gaussian,

$$p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) = \mathcal{N}(\boldsymbol{y}_t|\mathbf{H}_t\boldsymbol{\theta}_t, \mathbf{R}_t), \tag{72}$$

$$\hat{\boldsymbol{\theta}}^{(i)}_{t|t-1} \sim \mathcal{N}\left(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right). \tag{73}$$

Under these conditions the EnKF gain $\bar{\mathbf{K}}_t$ matches the KF gain $\mathbf{K}_t$, because

$$\mathrm{cov}_i\left[\hat{\boldsymbol{\theta}}^{(i)}_{t|t-1}, \hat{\boldsymbol{y}}^{(i)}_{t|t-1}\right] = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top, \tag{74}$$

$$\mathbb{V}_i\left[\hat{\boldsymbol{y}}^{(i)}_{t|t-1}\right] = \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top + \mathbf{R}_t, \tag{75}$$

$$\bar{\mathbf{K}}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top\left(\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top + \mathbf{R}_t\right)^{-1} \tag{76}$$

$$= \mathbf{K}_t, \tag{77}$$

and therefore the statistics of the posterior ensemble also match the KF update

$$\boldsymbol{\mu}_t = \mathbb{E}_i\left[\hat{\boldsymbol{\theta}}^{(i)}_t\right] = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\left(\boldsymbol{y}_t - \mathbf{H}_t\boldsymbol{\mu}_{t|t-1}\right), \tag{78}$$

$$\boldsymbol{\Sigma}_t = \mathbb{V}_i\left[\hat{\boldsymbol{\theta}}^{(i)}_t\right] = \boldsymbol{\Sigma}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}. \tag{79}$$

The weighted-likelihood EnKF: We propose a weighted version of EnKF based on our WoLF with dimension-specific weights and Gaussian likelihood, where (69) is modified to sample particle predictions from the weighted likelihood in (62):

$$\hat{\boldsymbol{y}}^{(i)}_{t|t-1} \sim \mathcal{N}\left(h_t\left(\hat{\boldsymbol{\theta}}^{(i)}_{t|t-1}\right), \bar{\mathbf{R}}_t\right). \tag{80}$$

The weighted EnKF converges to the WoLF as $N \to \infty$, by the same argument for the vanilla EnKF and EKF in (74) through (79). Specifically, under this limit the sampling scheme in (80) and the update in (70) yield

$$\bar{\mathbf{K}}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top\left(\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^\top + \bar{\mathbf{R}}_t\right)^{-1}, \tag{81}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \mathbf{H}_t\boldsymbol{\mu}_{t|t-1}\right), \tag{82}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t|t-1} - \bar{\mathbf{K}}_t\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}, \tag{83}$$

which matches the WoLF update in (66), (67), and (68).

For the case of scalar weights with $w_t^2 = W^2(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) \in \{0,1\}$ as in (19), the weighted EnKF can be implemented by unweighted sampling as in (69) followed by a weighted update that replaces (70) with

$$\hat{\boldsymbol{\theta}}^{(i)}_t = \hat{\boldsymbol{\theta}}^{(i)}_{t|t-1} + w_t^2\bar{\mathbf{K}}_t\left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}^{(i)}_{t|t-1}\right). \tag{84}$$

That is, the update is simply skipped when $w_t^2 = 0$, in agreement with Proposition 3.1 and Algorithm 1.

For the case of vector weights $\boldsymbol{w}_t$ as in Appendix D.2, when $\boldsymbol{w}_t \in \{0,1\}^d$ the weighted EnKF can be implemented by a generalisation of (84) that ignores only those observation components $\boldsymbol{y}_{t,j}$ for which $\boldsymbol{w}_{t,j} = 0$. Specifically, we can use the vanilla EnKF sampling in (69) but sample only the positively weighted dimensions, i.e. $\hat{\boldsymbol{y}}^{(i)}_{t|t-1,J_t}$ where $J_t = \{j : \boldsymbol{w}_{t,j} > 1\}$. Thus $\hat{\boldsymbol{y}}^{(i)}_{t|t-1}$ has size $|J_t|$ and (71) yields $\bar{\mathbf{K}}_t$ with size $m \times |J_t|$, and the particles can be updated according to

$$\hat{\boldsymbol{\theta}}^{(i)}_t = \hat{\boldsymbol{\theta}}^{(i)}_{t|t-1} + \bar{\mathbf{K}}_t\left(\boldsymbol{y}_{t,J_t} - \hat{\boldsymbol{y}}^{(i)}_{t|t-1,J_t}\right). \tag{85}$$

This agrees with the vanilla EnKF update except that zero-weighted observation dimensions are ignored. It can be shown to converge to the dimensionally weighted WoLF described in Appendix D.2 as $N \to \infty$.[9]

[9]The general case of $\boldsymbol{w}_t \in [0,1]^d$ can be implemented by combining this scheme with weighted sampling of positively weighted dimensions, i.e. using (80) to sample $\hat{\boldsymbol{y}}^{(i)}_{t|t-1,J_t}$.

For the experiments reported here (Section 4.3), we take the shortcut of sampling predictions for all dimensions and generalising (84) to vector weights:

$$\hat{\boldsymbol{\theta}}_t^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + \bar{\mathbf{K}}_t \mathrm{Diag}(\boldsymbol{w}_t) \left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)} \right). \tag{86}$$

This approximates the method of sampling only $\hat{\boldsymbol{y}}_{t|t-1,J_t}^{(i)}$ and updating with (85). This is more efficient in our Jax (Bradbury et al., 2018) implementation because all arrays are of constant size. We generalise (19) to define the weight vector as

$$\boldsymbol{w}_{t,j} = \begin{cases} 1 & \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{y}_{t,j} - \hat{\boldsymbol{y}}_{t|t-1,j}^{(i)} \right)^2 \leq c, \\ 0 & \text{otherwise}, \end{cases} \tag{87}$$

We call this the **average-particle EnKF (AP-EnKF)**.

Alternatively, we propose the update equation

$$\hat{\boldsymbol{\theta}}_t^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + \bar{\mathbf{K}}_t \, \mathrm{Diag} \left(\boldsymbol{w}_t^{(i)} \right) \left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)} \right), \tag{88}$$

with

$$\boldsymbol{w}_{t,j}^{(i)} = \begin{cases} 1 & \left(\boldsymbol{y}_{t,j} - \hat{\boldsymbol{y}}_{t|t-1,j}^{(i)} \right)^2 \leq c, \\ 0 & \text{otherwise}, \end{cases} \tag{89}$$

which we call the **per-particle EnKF (PP-EnKF)**.

The Huberised EnKF: The update equations for the PP-EnKF and AP-EnKF are related to the *Huberised* EnKF (H-EnKF) algorithm of Roh et al. (2013) that update particles according to

$$\hat{\boldsymbol{\theta}}_t^{(i)} = \hat{\boldsymbol{\theta}}_{t|t-1}^{(i)} + \bar{\mathbf{K}}_t \, \rho \left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}^{(i)} \right), \tag{90}$$

with

$$\rho(z) = \begin{cases} c & \text{if } z > c, \\ -c & \text{if } z < -c, \\ z & \text{otherwise}, \end{cases} \tag{91}$$

$c > 0$ a given threshold, and the function ρ is applied element-wise.

E. Additional numerical experriments

This section provides additional numerical experiments. In particular, in Appendix E.1, we discuss Figure 2 in more detail; in Appendix E.2, we conduct robustness checks for the numerical experiments presented in Section 4; and in Appendix E.3, we provide an additional experiment for online learning in a non-stationary environment with outlier measurements.

E.1. Description of the PIF

In this section, we discuss the 2d tracking problem of Section 4.1 in more detail. We generate $t = 20$ steps from (20) and the last measurement $\boldsymbol{y}_t$ is replaced with $\boldsymbol{y}_t^c = \boldsymbol{y}_t + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in [-5, 5] \times [-5, 5]$ is the outlier noise. Figure 9 (left) shows the weighting term for the WoLF methods as a function of the outlier noise $\boldsymbol{\epsilon}$. Alternatively, Figure 9 (right) shows the Mahalanobis distance between the prior predictive $\hat{\boldsymbol{y}}_t$ and the outlier measurement $\boldsymbol{y}_t^c$. We observe that the midpoint of the contours for all methods is not centred at $\boldsymbol{\epsilon} = (0, 0)$. This is because the weighting term is a function of the prior predictive and the measurement at time t. Hence, for the IMQ, the weighting is 1 only if the prior predictive equals the measurement, i.e., $\hat{\boldsymbol{y}}_t = \mathbf{H}_t \boldsymbol{\mu}_{t|t-1}$. This result explains the distored PIF observed in Figure 2.

Figure 10 shows the PIF for the WoLF-IMQ and the WoLF-TMD first shown in Figure 2. We observe that the PIF for the WoLF-TMD has a hard cutoff that strongly bounds the PIF values at the expense of higher PIF around a region near the cutoff boundary. In contrast, the WoLF-IMQ bounds the PIF values more softly and has lower maximum PIF than the WoLF-TMD.

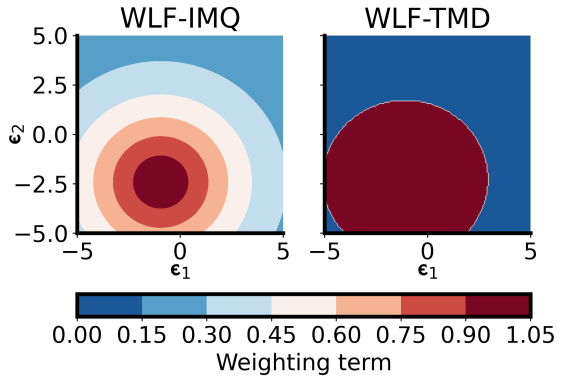

Figure 9. Weighting term $W_t(\boldsymbol{y}_{1:t})$ given an the outlier measurement $\boldsymbol{y}_t^c$. The x-axis and y-axis describe the errors $\boldsymbol{\epsilon} \in [-5, 5] \times [-5, 5]$.

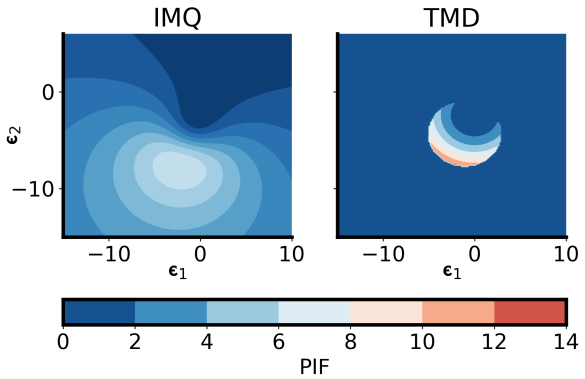

Figure 10. PIF of Figure 2 for extended domain of the errors $\boldsymbol{\epsilon} \in [-15, 10] \times [-15, 10]$.

E.2. Robustness

E.2.1. 2D TRACKING

In this section, we present further results from experiment 4.1. Figure 11 shows the distribution over errors over all 4 state components for both outlier variants of the 2d tracking problem. For the mixture variant, we see elongated tails in the error distribution for the **WoLF-TMD** algorithm. To see why, consider the trace of weights over time in Figure 12. We observe that both **WoLF-IMQ** and **WoLF-TMD** set the weighting term close to (or equal) to zero in outlier events. However, the **WoLF-TMD** is more prone to have false positives (at a rate of about 7% in this example), in which it fails to update the posterior state. This explains the elongated tails.

Right panel of Figure 3

2d tracking: hyperparameter choice Here we show the performance of each method for the experiment in Section 4.1 as we vary the hyperparameters. We do this for both, the Mixture case and the Student-t case.

Figure 14 shows the RMSE for estimating the first state component as we vary the c hyperparameter for the IMQ and the TDM weighting functions; for **KF-IW**, we use two inner iterations and we vary the h hyperparameter; for **KF-B**, we fix $\beta = 1.0$, use four inner iterations and vary the value of α. The definition for the RMSE $J_{T,0}$ is as in Section 4.1.

Empirical comparison to (Boustati et al., 2020) In this section, we compare the particle-filter-based method of (Boustati et al., 2020), which we denote the RBPF. Figure 15 extends Figure 4 to include the RBPF. The RBPF performs comparably to our proposed method; however, it has much higher computational cost and does not have a closed-form solution.

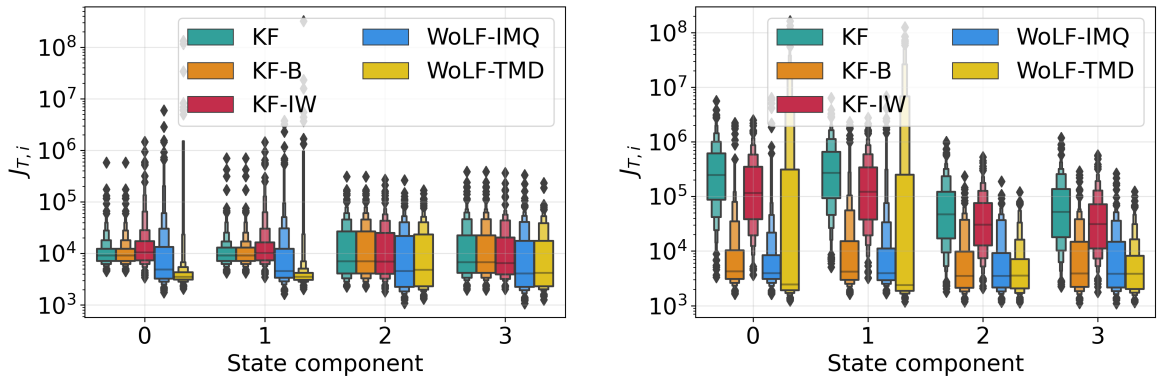

Figure 11. Distribution of $J_{T,i}$ for all state components. The left panel is for the student variant (21) and the right panel is for the mixture variant (22).

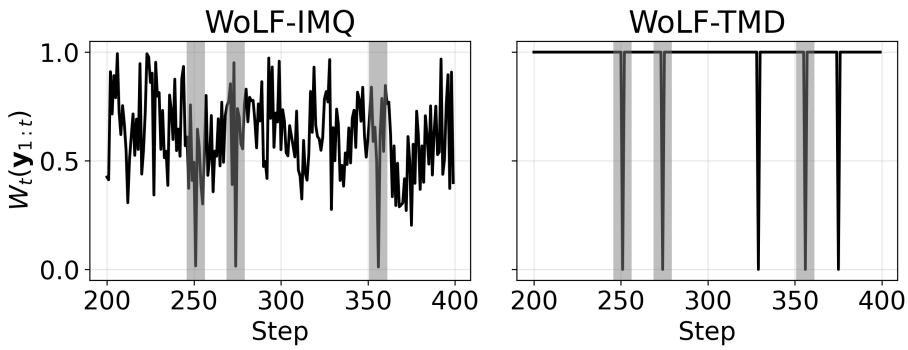

Figure 12. Trace of the weighting term for WoLF-IMQ and WoLF-MD.

E.2.2. UCI DATASETS

In this section, we give more details on the MLP regression experiments in Section 4.2. The size of the datasets and models which we use are summarised in Table 3.

Speed vs accuracy Figure 16 (left) shows the time-step and RMedSe for the Kin8nm dataset over multiple trials. Our methods, the **WoLF-IMQ** and the **WoLF-TMD**, strike the best balance between RMedSE and running time among the competing methods. The **EKF-IW** and the **EKF-B** have comparable error rates to the WoLF methods, but their running time is 6x and 12x slower than the **WoLF-IMQ**, respectively. The **EKF** has comparable running time, but 3x higher average RMedSE than the **WoLF-IMQ**. Finally, the **OGD** is 40% faster, but has 2x higher average RMedSE than the **WoLF-IMQ**.

Sensitivity to outlier rate We evaluate the sensitivity of the methods to the choice of corrosion rate p_ϵ for the Kin8nm dataset. Figure 16 (right) shows the RMedSE after 100 trials of each method as a function of p_ϵ. We set the hyperparameters using BO with $p_\epsilon = 0.1$ and evaluate the RMedSE with $p_\epsilon \in \{0.00, 0.05, \ldots, 0.45\}$. We observe that all robust methods — the **WoLF-IMQ**, the **WoLF-TMD**, the **EKF-B**, and the **EKF-IW**— have similar RMEdSE for any choice of p_ϵ and have similar RMedSE rate of increase. Conversely, the **EKF** and the **OGD** are much less stable and their RMedSE error rate increases at a faster pace.

E.2.3. LORENZ96 MODEL

In this subsection, we evaluate the EnKF methods in the Lorenz96 model presented in subsection 4.3 when we only have $N = 20$ particles but $d = 100$ states. In this scenario, the EnKF is usually modified to incorporate the *covariance inflation* proposed in Anderson (1999). We present the results for a single run of the modified methods in the left panel of Figure 17. In the right panel of Figure 17, we perform a sensitivity analysis to the choice of hyperparameter c. Similar to the results

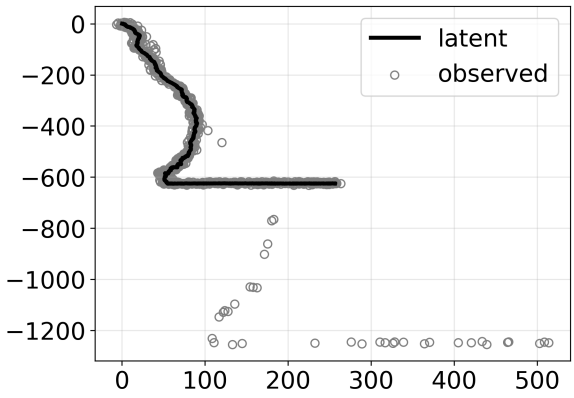

Figure 13. Sample of the top-left panel in Figure 3

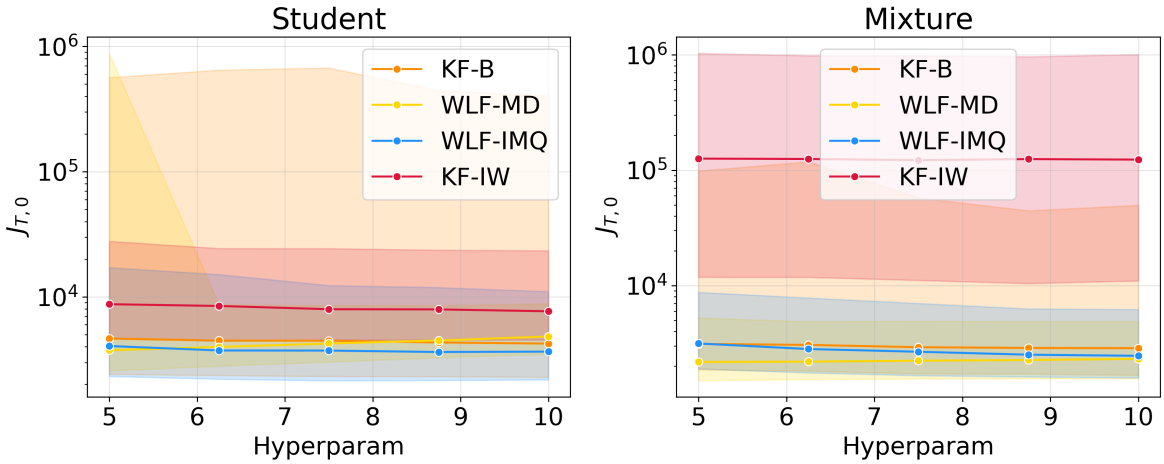

Figure 14. $J_{t,0}$ over 100 runs for the methods as a function of their hyperparameter.

presented in the main body of text, we observe that, in the best scenario, the **Hub−EnKF** and **AP−EnKF** behave similarly. However, the **Hub−EnKF** is more sensitive to the choice of hyperparameters than the **AP−EnKF** and **PP−EnKF**.

E.3. Robust EKF for online MLP regression (1d) — suplementary experiment

In this section, we consider an online nonlinear 1d regression, with the training data coming either from an i.i.d. source, or a correlated source. The latter corresponds to a non-stationary problem (see e.g. Cartea et al., 2023; Arroyo et al., 2024; Duran-Martin et al., 2022).

We present a stream of observations $\mathcal{D}^{\text{filter}} = (y_1, x_1), \ldots (y_T, x_T)$ with $y_t \in \mathbb{R}$ the measurements, $x_t \in \mathbb{R}$ the exogenous variables, and $T = 1500$. The measurements and exogenous variables are sequentially sampled from the processes

$$y_t = \begin{cases} \boldsymbol{\theta}_1^* x_t - \boldsymbol{\theta}_2^* \cos(\boldsymbol{\theta}_3^* x_t \, \pi) + \boldsymbol{\theta}_4^* x_t^3 + V_t & \text{w.p. } 1 - p_\epsilon, \\ U_t & \text{w.p. } \; p_\epsilon, \end{cases} \tag{92}$$

where the parameters of the observation model are $\boldsymbol{\theta}^* = (0.2, -10, 1.0, 1.0)$, the inputs are $x_t \sim \mathcal{U}[-3, 3]$, and the noise is $V_t \sim \mathcal{N}(0, 3)$, $U_t \sim \mathcal{U}[-40, 40]$, and $p_\epsilon = 0.05$.

We consider four configurations of this experiment. In each experiment the data is either sorted by x_t value (i.e, the exogenous variable satisfies $x_i < x_j$ for all $i < j$, representing a correlated source) or is unsorted (representing an i.i.d. source), and the measurement function is either a clean version of the true data generating process (i.e., (92) with $p_\epsilon = 0$ and unknown coefficients $\boldsymbol{\theta}$), or a neural network with unknown parameters $\boldsymbol{\theta}$. Specifically, we use a multi-layered perceptron

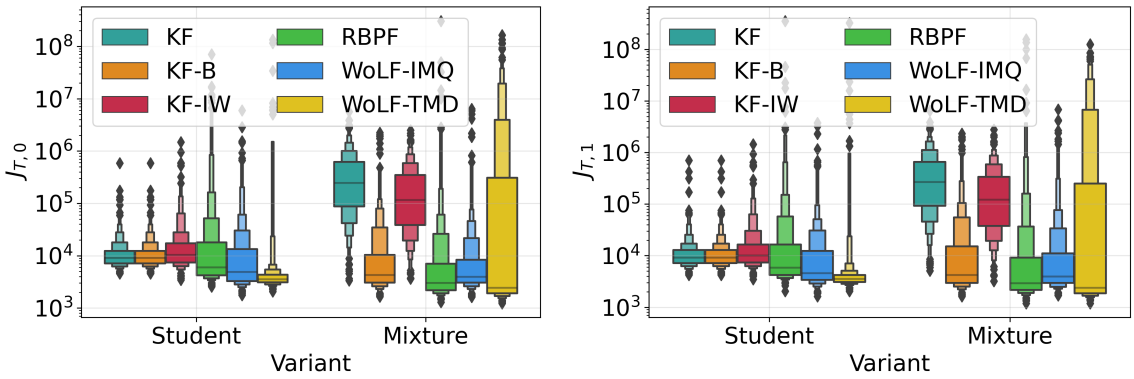

Figure 15. Distribution (across 500 2d tracking trials) of RMSE for first component of the state vector, $J_{T,0}$. Left panel: Student observation model. Right panel: Mixture observation model. We extend Figure 4 to include the `RBPF`.

Dataset	#Examples T	#Features d	#Parameters m
Boston	506	14	321
Concrete	1,030	9	221
Energy	768	9	221
Kin8nm	8,192	9	221
Naval	11,934	18	401
Power	9,568	5	141
Protein	45,730	10	241
Wine	1,599	12	281
Yacht	308	7	181

Table 3. Description of UCI datasets. Number of parameters refers to the size of the one-layer MLP.

(MLP) with two hidden layers and 10 units per layer:

$$h(\boldsymbol{\theta}_t, x_t) = \boldsymbol{w}_t^{(3)} \phi\left(\boldsymbol{w}_t^{(2)} \phi\left(\boldsymbol{w}_t^{(1)} x_t + \boldsymbol{b}_t^{(1)}\right) + \boldsymbol{b}_t^{(2)}\right) + \boldsymbol{b}_t^{(3)}, \tag{93}$$

with activation function $\phi(u) = \max\{0, u\}$ applied elementwise. Thus the state vector encodes the parameters:

$$\boldsymbol{\theta}_t = (\boldsymbol{w}_t^{(1)} \in \mathbb{R}^{10\times 1}, \boldsymbol{w}_t^{(2)} \in \mathbb{R}^{10\times 10}, \boldsymbol{w}_t^{(3)} \in \mathbb{R}^{1\times 10}, \boldsymbol{b}_t^{(1)} \in \mathbb{R}^{10}, \boldsymbol{b}_t^{(2)} \in \mathbb{R}^{10}, \boldsymbol{b}_t^{(3)} \in \mathbb{R})$$

and has size so that $\boldsymbol{\theta} \in \mathbb{R}^{141}$. Note that in this experiment $h_t(\boldsymbol{\theta}) = h(\boldsymbol{\theta}, x_t)$. We set $Q_t = 10^{-4}\mathbf{I}$, which allows the parameters to slowly drift over time and provides some reguralisation.

For each method, we evaluate the RMedSE $= \sqrt{\text{median}\{(y_t - h_t(\boldsymbol{\mu}_{t|t-1}))^2\}_{t=1}^T}$. The **`EKF-IW`** and the **`EKF-B`** methods are taken with two inner iterations, which implies that their computational complexity is twice that of the WoLF methods.

MLP measurement model Figure 18 shows results when the data are presented in sorted order of x_t. We show the performance on 100 trials. The left panel shows the mean prior-predictive $h(\boldsymbol{\mu}_{t|t-1}, x_t)$ of each method, and the underlying true state process, for a single trial. The right panel shows the RMedSE after multiple trials. We observe on the right panel that the **`WoLF-IMQ`** and the **`EKF-IW`** have the lowest mean error and lowest standard deviation among the competing methods. However, the **`EKF-IW`** takes twice as long to run the experiment. For all methods, the performance worse on the left-most side of the plot on the left panel, which is a region with not enough data to determine whether a measurement is an inlier or an outlier.

Figure 19 shows the results when data are presented in random order of $\boldsymbol{x}_t$. We show results for a single run on the left panel

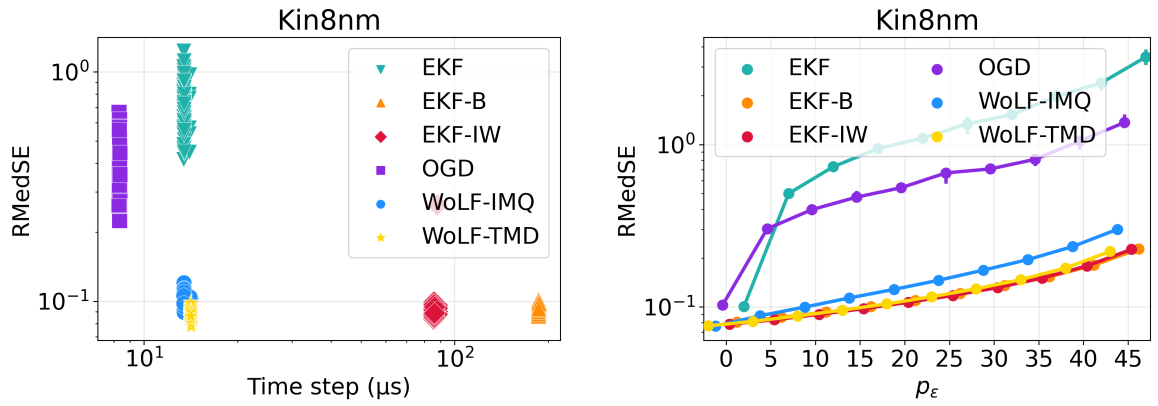

Figure 16. **Left panel** shows the RMedSE (root median squared error) vs running time (per observation) to fit a neural network to the Kin8nm UCI regression dataset. Each point corresponds to a different trial. **Right panel** shows the RMedSE (root median squared error) vs percentage probability of outlier, p_ϵ when fitting a neural network to the Kin8nm UCI regression dataset. (We show the bootstrp average over 100 trials and 95% confidence interval) Crucially, the same hyper-parameters are used for all experiments (and are estimated under $p_\epsilon = 0.1$).

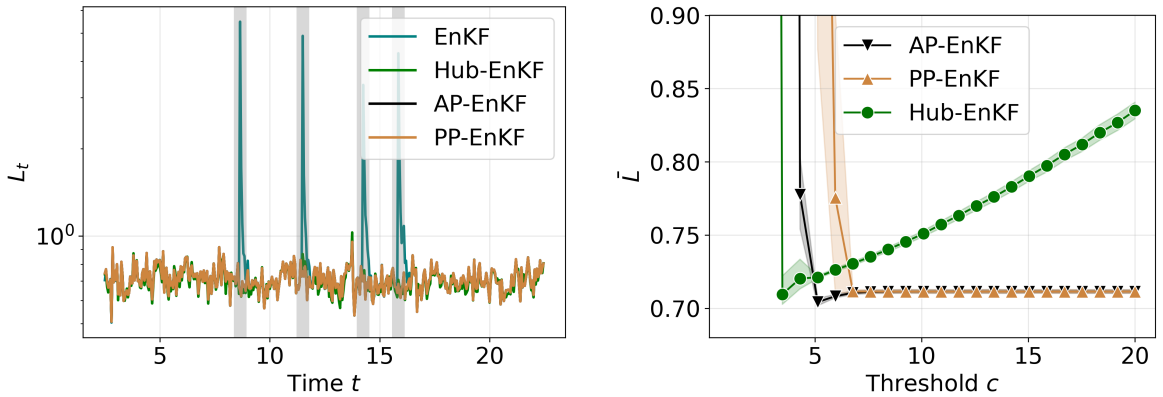

Figure 17. The left panel shows a run of the **EnKF**, the **AP-EnKF**, the **PP-EnKF**, and the **Hub-EnKF**; with covariance inflation; outlier events are shown in grey vertical bars. The right panel shows the bootstrap estimate of L_T over 20 runs and 500 bootstrapped samples as a function of the c hyperparameter.

and the the RMedSE after multiple trials on the right panel. Similar to the sorted configuration, we observe that the **EKF-IW** and the **WoLF-IMQ** are the methods with lowest RMedSE. However, the **EKF-IW** has longer tails than the **WoLF-IMQ**.

True measurement model We modify the experiment above by taking the measurement function to be $h_t(\boldsymbol{\theta}_t) = h(\boldsymbol{\theta}_t, x_t) = \boldsymbol{\theta}_{t,1} x_t - \boldsymbol{\theta}_{t,2} \cos(\boldsymbol{\theta}_{t,3} x_t \, \pi) + \boldsymbol{\theta}_{t,4} x_t^3$, with state $\boldsymbol{\theta}_t \in \mathbb{R}^4$ and $\boldsymbol{\theta}_{t,i}$ the i-th entry of the state vector $\boldsymbol{\theta}_t$. Figure 20 shows a single run of the filtering process when the data is presented unsorted (left panel) and sorted (right panel). We observe that the behaviour of the **WoLF-IMQ**, the **WoLF-TMD**, and the **EKF-IW** have similar performance. However, the **EKF-IW** takes twice the amount of time to run. The **OGD** and the **EKF** are not able to correctly filter out outlier measurement at the tails. Finally, the **EKF-B** over-penalises inliers and does not capture the curvature of the measurement process.

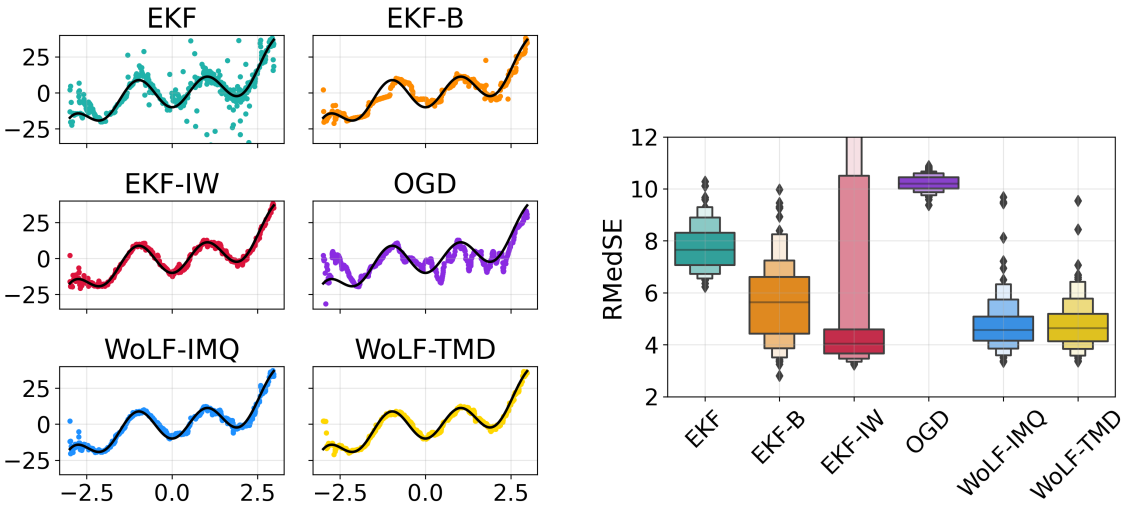

Figure 18. Results with sorted data. Left panel shows a run of each filter on the 1d regression, with the true underlying data-generating function in solid black line and the next-step predicted observation as dots. Right panel shows the RMedSE distribution over multiple trials.

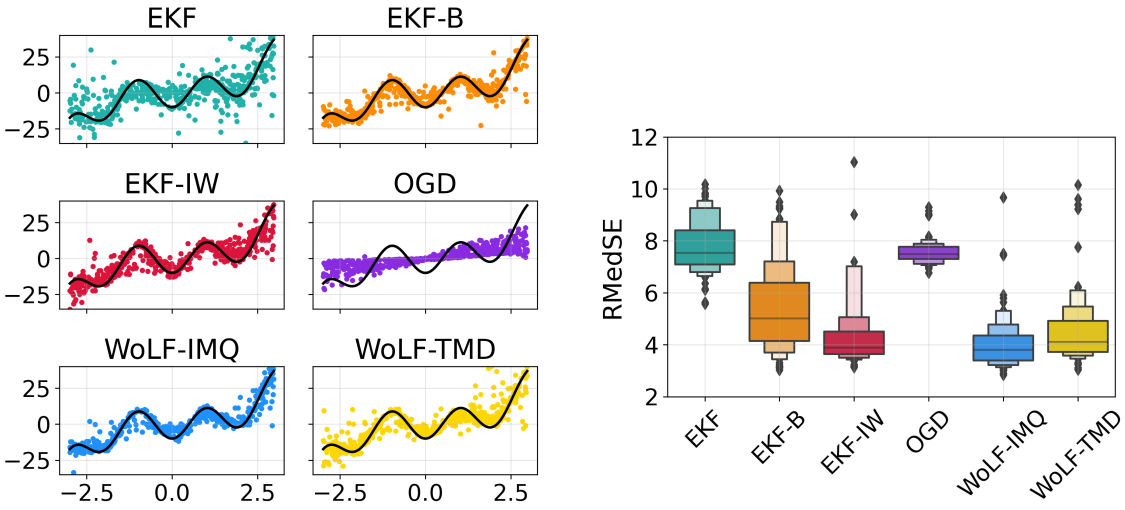

Figure 19. Results with unsorted inputs. The left panel shows a run of each filter with the underlying data-generating function in solid black line and the next-step predicted observation as dots. The right panel shows the distribution of G_T for multiple runs. We remove all values of G_T that have a value larger than 800.

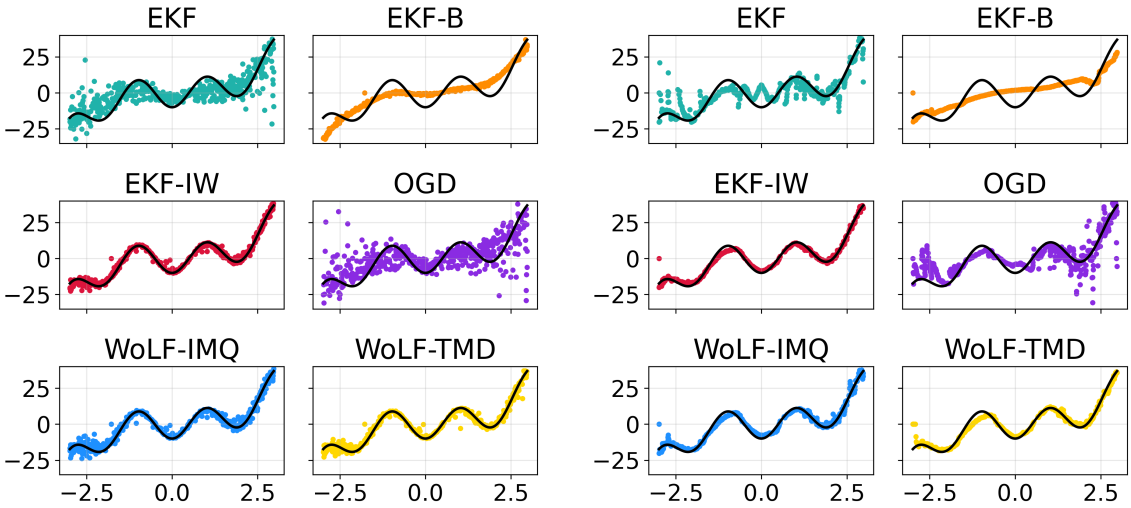

Figure 20. The figure shows a run of each filter with the underlying data-generating function in solid black line and the evaluation of $h(\boldsymbol{\mu}_{t|t-1}, x_t)$ in points. The left panel shows the configuration with unsorted x_t values and the right panel shows the configuration with sorted x_t values.