SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

April 2026, Vol. 69, Iss. 4, 142202:1–142202:16
https://doi.org/10.1007/s11432-024-4726-3

Adaptive-weighted variational Bayesian filter for robust state estimation without prior statistics

Juhui WEI[1], Bowen HOU[1], Dayi WANG[2], Wenchao XUE[3] & Jiongqi WANG[1*]

[1]*College of Science, National University of Defense Technology, Changsha 410072, China*
[2]*Beijing Institute of Spacecraft System Engineering, China Academy of Space Technology, Beijing 100191, China*
[3]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Received 6 August 2024/Revised 4 December 2024/Accepted 29 March 2025/Published online 15 January 2026

Abstract Non-Gaussian noise and measurement anomalies are major challenges in using filters for state estimation in engineering practice, particularly in the absence of relevant prior statistical information. Hence, we propose an adaptive-weighted variational Bayesian filter (AW-VBF) to obtain a robust state estimation. First, the variational Bayes approach is used to establish a fundamental filter framework that can handle both nonlinear and non-Gaussian scenarios by approximating the true posterior distribution through a parameterized distribution. Interestingly, the state transition and state measurement processes play independent roles in this framework. Subsequently, kernel density estimation is adopted to capture non-Gaussian noise characteristics from historical data. A novel adaptive weight function that is twice differentiable (thus ensuring the existence of gradient estimators) replaces the likelihood loss function to address measurement anomalies. The algorithm process, including the gradient estimator details, is provided. Target-tracking simulation results under different conditions verify the superiority of the AW-VBF to existing methods. Compared with these conventional methods, our method enhances position estimation accuracy by 37.96% and velocity estimation accuracy by 32.35% in the presence of non-Gaussian noise. The corresponding enhancements in the presence of measurement anomalies are 64.92% and 25.59%, respectively.

Keywords adaptive filter, anomalies, kernel density estimation, M-estimation, non-Gaussian noise

Citation Wei J H, Hou B W, Wang D Y, et al. Adaptive-weighted variational Bayesian filter for robust state estimation without prior statistics. Sci China Inf Sci, 2026, 69(4): 142202, https://doi.org/10.1007/s11432-024-4726-3

1 Introduction

State estimation provides fundamental data support for decision-making in various fields, including industries [1], aerospace [2], and energy [3]. Bayesian filters, the primary technology for state estimation, have been widely used in dynamic systems for reliable state estimation since their inception [4, 5]. For example, the classic Kalman filter (KF), a special Bayesian filter, has been theoretically proven to be optimal under linear and Gaussian conditions. Furthermore, several nonlinear Gaussian filters have been developed for complex systems, such as the extended KF (EKF) [6,7], unscented KF (UKF) [8], and cubature KF [9,10]. These filters use different Gaussian-weighted integral approximation criteria and can achieve at least second-order accuracy [11]. However, in engineering practice, in addition to the challenges entailed by nonlinearity, we are confronted with two other challenges arising from non-Gaussian noise and measurement anomalies, thereby making it difficult for us to obtain a stable state estimation result. In recent years, researchers have attempted to address the aforementioned challenges for more reliable and robust state estimation.

Non-Gaussian noise might originate from modeling approximations or the variability of working environments. When noise characteristics (uncalibrated or time varying) substantially deviate from the preset values, they will affect the filter performance [12]. One of the popular technologies is based on information theoretic learning [13]. For example, the Rényi entropy [14] and correntropy [15] are utilized as adaptation criteria to improve the robustness of state estimation in the presence of heavy-tailed impulsive noise. Furthermore, several information-based algorithms perform better under complex noise distributions; examples include Gaussian entropy [16], generalized correntropy [17], least stochastic entropy [18], centered error entropy [19], and other approaches [20–23]. Using a particle filter (PF) based on Monte-Carlo sampling can estimate states in strongly nonlinear and non-Gaussian systems without assuming the noise distribution type [24]; however, its application is limited by degeneracy [25, 26].

* Corresponding author (email: wjq_gfkd@163.com)

Measurement anomalies are frequently induced by environmental interference or unpredictable sensor malfunctions, which are assumed to exert a sustained adverse impact on system state estimation. Scalable approaches are applied to filters by detecting measurements that deviate significantly from the data center. Ref. [27] developed a robust filter based on a detect-and-reject idea; this filter automatically identifies anomalies by iteratively using a mean-field variational Bayesian method at each time instant. Ref. [28] modeled measurement anomalies from an impulsive signal and developed a parameter-dependent set-membership filter to determine whether the current measurement output is abnormal. Ref. [29] unified randomly occurring measurement anomalies, modeled them as a Dirichlet distribution, and simultaneously estimated the state and measurement anomaly parameters through a robust variational filter. Some other robust general techniques have been derived from M-estimation statistics. Using robust loss functions, these filters can effectively minimize the interference of anomalies and avoid large estimation errors [30]. In addition, similar methods have been used in engineering practice [31, 32].

Although scholars have made various improvements based on classical filters to adapt to non-Gaussian noise and measurement anomalies, the superiority of their approaches inevitably depends on practical applications. The variational Bayesian filter (VBF), a novel framework, selects the closest result to the true posterior probability distribution from a specified parameterized distribution cluster. Ref. [33] originally applied variational Bayes to joint recursive estimation, proposed the variational-Bayesian-based adaptive KF algorithm for time varying measurement noise, and verified its feasibility through simulations. Various nonlinear adaptive VBFs have since been proposed to solve problems involving unknown measurement noise [34, 35]. Refs. [36–38] modeled a predicted covariance matrix using an inverse Wishart distribution for robust state estimation in a VBF. Ref. [39] embedded maximum likelihood (ML) estimation into a VBF and applied the strong robust centered error entropy criterion to handle measurement noise. However, the complex computation of this VBF limits its further development. Ref. [40] proposed a classic solution based on the mean-field assumption to seek an approximate distribution close to the target joint. Ref. [41] presented an alternative algorithm based on stochastic optimization, which allows for the direct optimization of the variational lower bound and avoids nonclosed-form integrals in the mean-field method. Ref. [42] applied a natural gradient method to the Kullback-Leibler (KL) divergence and obtained a closed-form iterative procedure of variational parameters. In addition, various engineering applications of VBF have been attempted [43, 44].

In summary, although existing filters (including VBF) have been specifically improved, they are mostly dependent on prior information. In fact, previous studies frequently assumed a specific form for non-Gaussian noise and derived the corresponding loss functions for dealing with it. When handling measurement anomalies, they typically detect them based on prior information and assign the weights of 0 and 1 to eliminate the influence of anomalies.

Therefore, we present some innovative approaches in this paper for dealing with non-Gaussian noise and measurement anomalies without relying on prior statistics. The detailed contributions of this paper are as follows.

(1) The individual roles of state transfer and state measurement processes in the VBF are identified, and improvements can be made according to requirements.

(2) Kernel density estimation (KDE) is used to approximate non-Gaussian process noise.

(3) A novel weighting function that is twice differentiable (to ensure the existence of gradient estimators) is designed to handle measurement anomalies.

(4) The adaptive-weighted VBF (AW-VBF) is proposed; the algorithm process, including the details of the gradient estimator, is provided.

The rest of this article is organized as follows. In Section 2, we introduce the issue of state estimation and expound the forms and potential influences of non-Gaussian noise and measurement anomalies. In Section 3, AW-VBF is proposed to improve the robustness of state estimation. The details of this algorithm are also presented in this section. Section 4 contains the simulation details and a systematic analysis of the performance of our method and comparative methods.

2 Problem formulation

Consider the following general nonlinear discrete dynamical system with nonlinear measurements as

$$\begin{cases} x_k = f_{k|k-1}\left(x_{k-1}\right) + w_{k-1}, \\ y_k = h_k\left(x_k\right) + v_k, \end{cases} \tag{1}$$

where $x_k \in \mathbb{R}^{n_x}$ and $y_k \in \mathbb{R}^{n_y}$ are the system state and measurement, respectively, at time t_k; $f_{k|k-1}(\cdot)$ and $h_k(\cdot)$ are the state transfer function and measurement function, respectively; w_{k-1} and v_k are the state noise and

measurement noise, respectively. Usually, system (1) meets the Markov condition

$$\begin{cases} p\left(x_k|x_{1:k-1}, y_{1:k-1}\right) = p\left(x_k|x_{k-1}\right), \\ p\left(x_{k-1}|x_{k:k+T}, y_{k:k+T}\right) = p\left(x_{k-1}|x_k\right), \\ p\left(y_k|x_{1:k}, y_{1:k-1}\right) = p\left(y_k|x_k\right), \end{cases} \tag{2}$$

where $x_{1:k} = \{x_1, x_2, \ldots, x_k\}$ and $y_{1:k} = \{y_1, y_2, \ldots, y_k\}$.

A filter aims to obtain the current system state estimation $\hat{x}_k$ using the current and historical observation data $y_{1:k}$. Then, the optimal nonlinear filtering problem aims to find the posterior distribution $p\left(x_k|y_{1:k}\right)$ of x_k with the given measurement y_k. According to the Bayesian principle, the posterior distribution of the system state is

$$p\left(x_k|y_{1:k}\right) = \frac{p\left(y_k|x_k\right) p\left(x_k|y_{1:k-1}\right)}{p\left(y_k|y_{1:k-1}\right)}, \tag{3}$$

where

$$\begin{cases} p\left(x_k|y_{1:k-1}\right) = \mathbb{E}_{p(x_{k-1}|y_{1:k-1})}\left[p\left(x_k|x_{k-1}\right)\right] = \displaystyle\int p\left(x_k|x_{k-1}\right) p\left(x_{k-1}|y_{1:k-1}\right) dx_{k-1}, \\ p\left(y_k|y_{1:k-1}\right) = \mathbb{E}_{p(x_k|y_{1:k-1})}\left[p\left(y_k|x_k\right)\right] = \displaystyle\int p\left(y_k|x_k\right) p\left(x_k|y_{1:k-1}\right) dx_k. \end{cases} \tag{4}$$

The integration in (4) is difficult, except in the case of linear Gaussian systems (in Appendix A).

For recursive computations, we assume that the state estimate $\hat{x}_{k-1}$ and its posterior distribution $p\left(x_{k-1}|y_{1:k-1}\right)$ at time t_{k-1} are known. System (1) is reasonably rewritten into the form

$$\tilde{y}_k = \tilde{h}_k\left(x_k\right) + \tilde{v}_k, \tag{5}$$

where

$$\tilde{y}_k = \begin{bmatrix} f_{k|k-1}\left(\hat{x}_{k-1}\right) \\ y_k \end{bmatrix}, \quad \tilde{h}_k\left(x_k\right) = \begin{bmatrix} x_k \\ h_k\left(x_k\right) \end{bmatrix}, \quad \tilde{v}_k = \begin{bmatrix} -\hat{w}_k \\ v_k \end{bmatrix}. \tag{6}$$

In accordance with (3), the ML estimation of the system state x_k is

$$\left(\hat{x}_k\right)_{ML} = \arg\max_{\hat{x}_k} p\left(\tilde{y}_k \middle| \tilde{h}_k\left(\hat{x}_k\right)\right) = \arg\max_{\hat{x}_k} \log p\left(\tilde{y}_k \middle| \tilde{h}_k\left(\hat{x}_k\right)\right). \tag{7}$$

If $\tilde{v}_k$ meets the Gaussian noise, then Eq. (7) can degenerate into generalized least squares estimation:

$$\left(\hat{x}_k\right)_{GLS} = \arg\min_{\hat{x}_k} e_k^{\mathrm{T}} \tilde{R}_k^{-1} e_k, \quad e_k = \tilde{y}_k - \tilde{h}_k\left(\hat{x}_k\right), \tag{8}$$

where $\tilde{R}_k$ is the covariance matrix of the noise $\tilde{v}_k$.

However, non-Gaussian noise and measurement anomalies pose substantial challenges in this method, especially when their statistical properties remain unknown. Here, non-Gaussian noise undermines the quadratic form of the loss function as shown in (8), resulting in a disparity between the least squares and ML estimations. Concurrently, non-Gaussian noise also gives rise to computational difficulties in (4). In addition, measurement anomalies refer to measurements that substantially deviate from the data center, typically signifying $|y_k - h_k\left(x_k\right)| > 3\sigma$. These abnormal measurements will induce a wrong estimation result because their occurrence probability is nearly zero.

M-estimation, as a robust estimation theory, can maintain its performance in non-Gaussian process noise or abnormal measurements. Then, the M-estimate of the system state x_k is as follows:

$$\left(\hat{x}_k\right)_M = \arg\min_{\hat{x}_k} \sum_{i=1}^{n_x+n_y} \rho\left(r_k^{(i)}\right), \quad r_k = \hat{R}_k^{-1/2} \cdot e_k, \tag{9}$$

where $\rho\left(\cdot\right)$ is the robust M-estimate objective function, which is usually a real-valued, even function. $\hat{R}_k$ is the robust estimation of residual covariance, which can be expressed as

$$\hat{R}_k = \left\{\hat{s}_{ij}\right\}, \quad \hat{s}_{ij} = \frac{1}{L-1}\left(e_k^{(i)} - \mathrm{med}\left[e_{k+1-L:k}^{(j)}\right]\right) \cdot \left(e_k^{(j)} - \mathrm{med}\left[e_{k+1-L:k}^{(j)}\right]\right), \tag{10}$$

where $\mathrm{med}\left(\cdot\right)$ is the median operator over the data window $e_{k+1-L:k}^{(j)}$.

Furthermore, M-estimation (9) can be written in the weighted form

$$(\hat{x}_k)_W = \arg\min_{\hat{x}_k} \sum_{i=1}^{n_x+n_y} \varphi\left(r_k^{(i),2}\right) \cdot r_k^{(i),2}, \quad r_k^{(i),2} = \left(r_k^{(i)}\right)^2, \tag{11}$$

where $\varphi(\cdot)$ is an adaptive weight function. In this equivalent form, only different weighting functions are required to be designed, such as those by Andrew [45], Hampel [46], and Huber [47].

Up to now, we have obtained a distinct understanding of this problem and the fundamental concepts of M-estimation. In the subsequent sections, we will learn the incorporation of these concepts into (3) and (5).

3 Proposed AW-VBF algorithm

3.1 VBF

According to the variational Bayesian principle, a parameterized posterior distribution $q_{\phi_k}(x_k|y_{1:k})$ is introduced to approximate the unknown true posterior distribution $p(x_k|y_{1:k})$. Unlike the approximate posterior in mean-field variational inference, it is not necessarily factorial, and its parameters ϕ_k are not computed from some closed-form expectation.

The marginal likelihood of individual data points y_k is rewritten as

$$\begin{aligned}\log p(y_k|y_{1:k-1}) &= \mathbb{E}_{q_{\phi_k}(x_k|y_{1:k})}\left[\log \frac{p(y_k|x_k)\, p(x_k|y_{1:k-1})}{p(x_k|y_{1:k})}\right] \\ &= \mathbb{E}_{q_{\phi_k}(x_k|y_{1:k})}\left[\log \frac{q_{\phi_k}(x_k|y_{1:k})\, p(y_k, x_k|y_{1:k-1})}{p(x_k|y_{1:k})\, q_{\phi_k}(x_k|y_{1:k})}\right] \\ &= \mathbb{D}_{KL}\left(q_{\phi_k}(x_k|y_{1:k})|p(x_k|y_{1:k})\right) + \mathcal{L}(\phi_k; y_k).\end{aligned} \tag{12}$$

The first right-hand side (RHS) term is the KL divergence (a nonnegative term) of the approximate posterior and true posterior. The second RHS term $\mathcal{L}(\phi_k; y_k)$ is called the (variational) lower bound on the marginal likelihood of the data point y_k and is written as follows:

$$\begin{aligned}\mathcal{L}(\phi_k; y_k) &= \mathbb{E}_{q_{\phi_k}(x_k|y_{1:k})}\left[\log p(y_k, x_k|y_{1:k-1}) - \log q_{\phi_k}(x_k|y_{1:k})\right] \\ &= \mathbb{E}_{q_{\phi_k}(x_k|y_{1:k})}\left[\log p(y_k|x_k) + \log p(x_k|y_{1:k-1}) - \log q_{\phi_k}(x_k|y_{1:k})\right] \\ &= \mathbb{E}_{q_{\phi_k}(x_k|y_{1:k})}\left[\log p(y_k|x_k)\right] - \mathbb{D}_{KL}\left(q_{\phi_k}(x_k|y_{1:k})|p(x_k|y_{1:k-1})\right).\end{aligned} \tag{13}$$

The purpose of the VBF is to find the optimal parameter ϕ_k that makes the approximate posterior $q_{\phi_k}(x_k|y_{1:k})$ and true posterior $p(x_k|y_{1:k})$ as close as possible, which means

$$\phi_k^* = \arg\min_{\phi_k} \mathbb{D}_{KL}\left(q_{\phi_k}(x_k|y_{1:k})|p(x_k|y_{1:k})\right). \tag{14}$$

Because the marginal likelihood $\log p(y_k|y_{1:k-1})$ is deterministic, Eq. (14) is equivalent to the optimization

$$\phi_k^* = \arg\max_{\phi_k} \mathcal{L}(\phi_k; y_k) = \arg\max_{\phi_k} \Big(\mathbb{E}_{q_{\phi_k}}\left[\log p(y_k|x_k)\right] - \mathbb{D}_{KL}\left(q_{\phi_k}|p(x_k|y_{1:k-1})\right)\Big). \tag{15}$$

In (15), the parameterized posterior distribution and prior information are indispensable within the variational Bayesian filter. The parameterized posterior distribution can be directly selected from the exponential family of distributions, as the mean and variance of the state estimation remain under our observation. Therefore, we set $\phi_k = \{\hat{x}_k, P_k\}$ and assume that $q_{\phi_k}(x_k|y_{1:k})$ follows a normal distribution, which means

$$q_{\phi_k}(x_k|y_{1:k}) = \mathcal{N}(x_k; \hat{x}_k, P_k). \tag{16}$$

Then, Eq. (15) can be calculated if the prior information is known, such as in system (A1). However, prior information is typically unknown because of the presence of non-Gaussian noise and measurement anomalies, constraining the application of VBF.

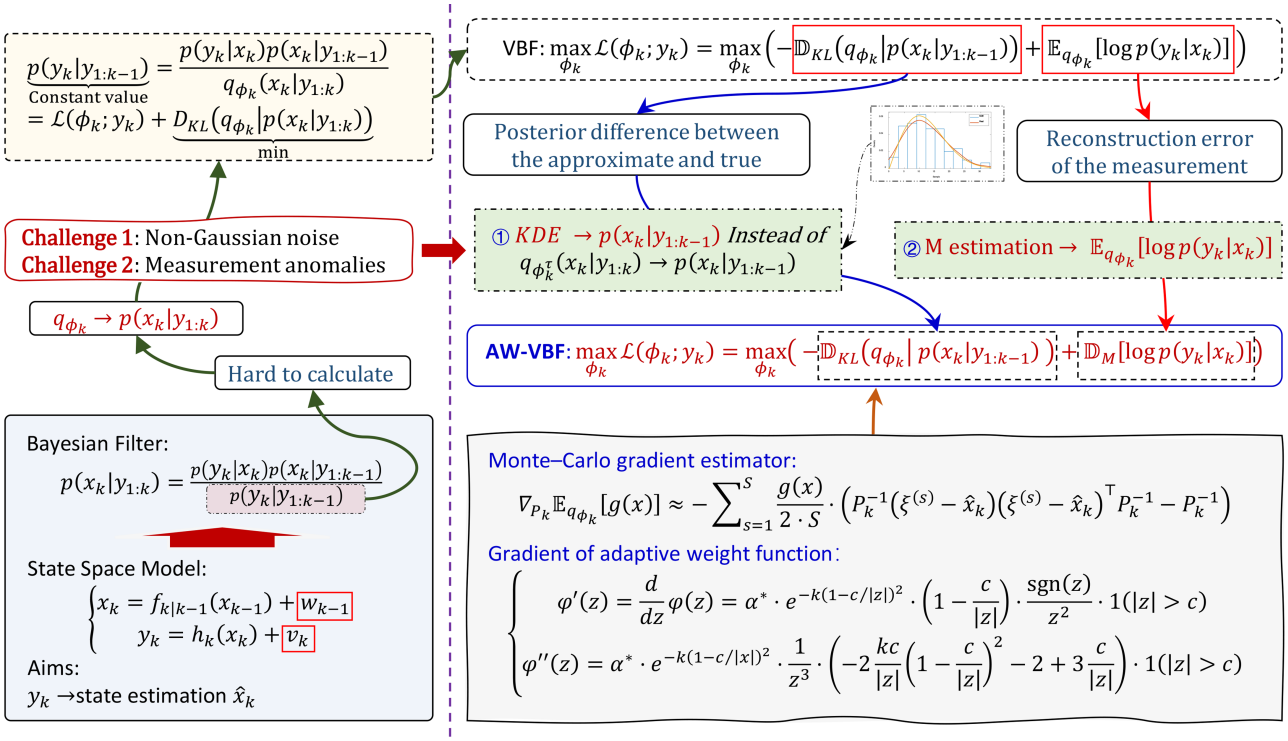

Figure 1 (Color online) Schematic diagram of motivation and algorithmic framework.

Specific example with the linear Gaussian assumption. We apply VBF to the linear Gaussian system (A1). Simple calculation (Appendix B) shows that

$$\mathcal{L}\left(\phi_k; y_k\right) = \frac{1}{2}\left(\begin{array}{r} C_0 - \left(\hat{x}_k - F\hat{x}_{k-1}\right)^{\mathrm{T}} P_{k|k-1}^{-1}\left(\hat{x}_k - F\hat{x}_{k-1}\right) - \left(y_k - H\hat{x}_k\right)^{\mathrm{T}} R_k^{-1}\left(y_k - H\hat{x}_k\right) \\ + \log\left|P_k\right| - \mathrm{Tr}\left(R_k^{-1} H P_k H^{\mathrm{T}}\right) - \mathrm{Tr}\left(P_{k|k-1}^{-1} P_k\right) \end{array}\right), \tag{17}$$

where $C_0 = n_x - n_y \log\left(2\pi\right) - \log\left|R_k\right| - \log\left|P_{k|k-1}\right|$. Considering the derivative of (17) with respect to ϕ_k yields

$$\begin{cases} \hat{x}_k = \left(P_{k|k-1}^{-1} + H^{\mathrm{T}} R_k^{-1} H\right)^{-1}\left(H^{\mathrm{T}} R_k^{-1} y_k + P_{k|k-1}^{-1} F\hat{x}_{k-1}\right), \\ P_k^{-1} = H^{\mathrm{T}} R_k^{-1} H + P_{k|k-1}^{-1}. \end{cases} \tag{18}$$

This easily verifies that Eq. (18) is consistent with KF (A4).

3.2 Motivation and algorithm framework

Considering the issue that the prior information remains unknown, we have conducted a detailed analysis of the VBF mechanism and have discovered an interesting fact that the state transition and measurement processes operate independently in this filter. Motivated by this finding, we proposed an AW-VBF without prior statistic information, as depicted in Figure 1. In this filter, KDE is used to handle non-Gaussian noise, and an adaptive weight function is adopted to deal with anomalies.

In-depth examination of variational lower bound. The first RHS term of (13) is written as

$$\mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|x_k\right)\right] = \int q_{\phi_k}\left(x_k|y_{1:k}\right) \cdot \log p\left(y_k|x_k\right) dx_k. \tag{19}$$

The second RHS term of (13) is written as

$$\mathbb{D}_{KL}\left(q_{\phi_k}|p\left(x_k|y_{1:k-1}\right)\right) = \int q_{\phi_k}\left(x_k|y_{1:k}\right) \cdot \log \frac{q_{\phi_k}\left(x_k|y_{1:k}\right)}{\int p\left(x_k|x_{k-1}\right) p\left(x_{k-1}|y_{1:k-1}\right) dx_{k-1}} dx_k. \tag{20}$$

Interestingly, Eq. (19) is completely determined by the state measurement process, whereas Eq. (20) is completely determined by the state transition process. Thus, the first RHS term of (13) can be interpreted as the reconstruction

error of the measurement, and its second RHS term can be deemed the difference between the approximate posterior and the true posterior derived from the state transition.

KDE for Non-Gaussian noise. As a nonparametric method, KDE can estimate distributions using historical data without assuming the distribution types. Experience has shown that different kernel functions can generate similar estimation results. In order to obtain clearer and more formal conclusions, a Gaussian kernel is adopted, which is expressed as

$$\kappa\left(z;\Lambda\right)=\left(2\pi\right)^{-n_z/2}\left|\Lambda\right|^{-1/2}\mathrm{e}^{-\left(z^{\mathrm{T}}\Lambda^{-1}z\right)/2}. \tag{21}$$

Therefore, given the historical state estimation $\hat{x}_i\,(i=k-L-1,k-L,\ldots,k-1)$, unknown and non-Gaussian $p\left(x_k|x_{k-1}\right)$ and $p\left(x_y|x_k\right)$ can be obtained using KDE as follows:

$$\begin{cases} p\left(x_k|x_{k-1}\right)=\dfrac{1}{L}\displaystyle\sum_{i=k-L}^{k-1}\kappa\left(x_{k-1}-\left(f_{k|k-1}\left(x_{k-1}\right)+\hat{x}_i-f_{i|i-1}\left(\hat{x}_{i-1}\right)\right);\Lambda_x^{(i)}\right), \\ p\left(y_k|x_k\right)=\dfrac{1}{L}\displaystyle\sum_{i=k-L}^{k-1}\kappa\left(y_k-\left(h_k\left(x_k\right)+y_i-h_i\left(\hat{x}_i\right)\right);\Lambda_y^{(i)}\right), \end{cases} \tag{22}$$

where $\Lambda_x^{(i)}$ and $\Lambda_y^{(i)}$ are the scale matrices of the state transition process and state measurement process, respectively, at time t_i.

Moreover, by replacing $p\left(x_{k-1}|y_{1:k-1}\right)$ with $q_{\phi_{k-1}}\left(x_{k-1}|y_{1:k-1}\right)$ and specifying $\Lambda=\lambda\cdot I$ [48], we obtain

$$\begin{aligned} p\left(x_k|y_{1:k-1}\right)&\approx\int p\left(x_k|x_{k-1}\right)q_{\phi_{k-1}}\left(x_{k-1}|y_{1:k-1}\right)dx_{k-1} \\ &=\frac{1}{L}\sum_{i=k-L}^{k-1}\mathbb{E}_{q_{\phi_{k-1}}}\left[\mathcal{N}\left(x_k;f_{k|k-1}\left(x_{k-1}\right)+\left(\hat{x}_i-f_{i|i-1}\left(\hat{x}_{i-1}\right)\right),\lambda\cdot I\right)\right]. \end{aligned} \tag{23}$$

Combining (19), (20), (22) and (23) yields

$$\begin{cases} \mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|x_k\right)\right]=C_1+\mathbb{E}_{q_{\phi_k}}\left[\log\displaystyle\sum_{i=k-L}^{k-1}\mathrm{e}^{-\frac{\left(y_k-\mu_y^{(i)}\right)^{\mathrm{T}}\left(y_k-\mu_y^{(i)}\right)}{2\lambda}}\right], \\ \mathbb{D}_{KL}\left(q_{\phi_k}|p\left(x_k|y_{1:k-1}\right)\right)=C_2-\dfrac{\log|P_k|}{2}-\mathbb{E}_{q_{\phi_k}}\left[\log\displaystyle\sum_{i=k-L}^{k-1}\mathbb{E}_{q_{\phi_{k-1}}}\left[\mathrm{e}^{-\frac{\left(x_k-\mu_x^{(i)}\right)^{\mathrm{T}}\left(x_k-\mu_x^{(i)}\right)}{2\lambda}}\right]\right], \end{cases} \tag{24}$$

where C_1 and C_2 are constant values and

$$\mu_x^{(i)}=f_{k|k-1}\left(x_{k-1}\right)+\hat{x}_i-f_{i|i-1}\left(\hat{x}_{i-1}\right),\quad \mu_y^{(i)}=h_k\left(x_k\right)+y_i-h_i\left(\hat{x}_i\right). \tag{25}$$

Adaptive weight function for measurement anomalies. Although KDE can effectively deal with non-Gaussian noise, it cannot handle anomalies in measurements, as these anomalies markedly deviate from historical data and are exceedingly rare in statistical terms.

Eqs. (19) and (24) characterize the reconstruction error using the marginal likelihood. In fact, considering the simple case $p\left(y_k|x_k\right)\sim\mathcal{N}\left(0,R_k\right)$, Eq. (19) can be expressed as

$$\mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|x_k\right)\right]=C_1-\frac{1}{2}\mathbb{E}_{q_{\phi_k}}\left[\left(y_k-h_k\left(x_k\right)\right)^{\mathrm{T}}R_k^{-1}\left(y_k-h_k\left(x_k\right)\right)\right]. \tag{26}$$

Inspired by the M-estimation method, an appropriate weight function can be devised for the reconstruction error to replace (26), that is

$$\mathbb{D}_M\left(\phi_k;y_k\right)=-\frac{1}{2}\mathbb{E}_{q_{\phi_k}}\left[\sum_{i=1}^{n_y}\varphi\left(r_k^{(i),2}\right)\cdot r_k^{(i),2}\right],\quad r_k=\hat{R}_k^{-1/2}\left(y_k-h_k\left(x_k\right)\right). \tag{27}$$

Previously reported weight functions, which are constant (0 and 1), do not downweight large residuals sufficiently, or are nondifferentiable. To address these issues and ensure that Eq. (27) demonstrates robustness against anomalies, we should assign equal weights to all smaller residuals (not greater than a cutoff value) and exponentially

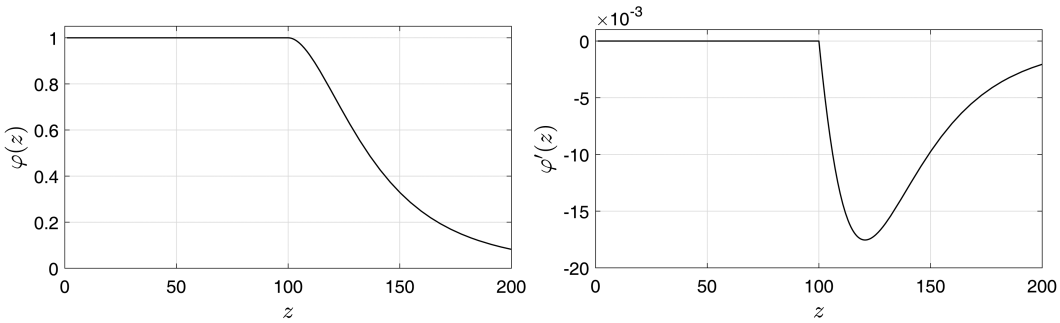

Figure 2 The graph of the weight function $\varphi(z)$, where $k = 10$ and $c = 100$ are chosen. The left subfigure depicts the values of the weight function, and the right subfigure presents its derivative.

downweight (penalize) larger residuals (greater than the cutoff value). Here, we consider a novel weight function [49], as an ad hoc choice, in the form of

$$\varphi(z) = \mathbb{I}(|z| \leqslant c) + \mathbb{I}(|z| > c) \frac{\mathrm{e}^{-k(1-c/|z|)^2} - \mathrm{e}^{-k}}{1 - \mathrm{e}^{-k}}, \tag{28}$$

where the tuning parameter $k > 1$ is a positive number (usually between 1 and 10) controlling the steepness of the exponentially decreasing weight, as depicted in Figure 2. The larger the k, the steeper the curve. The tuning parameter c is the point where the weight function changes from a constant one to an exponentially decreasing one. It is usually set to a large positive number, or it can be residual dependent (for example, the 50% or 75% percentile of residuals); a larger c indicates higher efficiency.

For the novel weight function $\varphi(z)$, verifying the following is straightforward:

(1) Function $\varphi(z)$ is twice differentiable, and $0 < \varphi(z) \leqslant 1$;

(2) When $z \to \infty$, $\varphi(z)$ is asymptotically equivalent to $\alpha(\mathrm{e}^{\gamma/z} - 1)$ for positive constants α and γ;

(3) If $z \to \infty$, then $\varphi(z)\, z \to 2ck/(\mathrm{e}^k - 1)$.

Notably, the novel weight function is twice differentiable, enabling its optimization in the subsequent AW-VBF solution without the concern of its gradient problem. Furthermore, we assign a weight one to residuals that lie close to the data center; points on the outskirts of the data cloud can be viewed as anomalies, so a lower positive weight should be given.

3.3 AW-VBF algorithm and gradient estimator

In Section 3.2, we proposed KDE for non-Gaussian noise and adaptive weight function for measurement anomalies. Combining (24) and (27), the optimization (15) can be written as

$$\phi_k^* = \arg\max_{\phi_k} \left(\mathbb{D}_M(\phi_k; y_k) - \mathbb{D}_{KL}\left(q_{\phi_k}(x_k|y_{1:k}) | p(x_k|y_{1:k-1})\right)\right). \tag{29}$$

Although Eqs. (17) and (18) provide the results of VBF with the linear Gaussian assumption, the universal analytical expression of AW-VBF is almost impossible to obtain.

Therefore, an iterative approach is required to solve the optimization (29), and considering the Taylor expansion of $\mathcal{L}(\phi_k; y_k)$ with respect to the variational parameter ϕ_k^τ,

$$\mathcal{L}(\phi_k; y_k) = \mathcal{L}(\phi_k^\tau; y_k) + \nabla_{\phi_k}\mathcal{L}(\phi_k; y_k)\big|_{\phi_k=\phi_k^\tau} \Delta\phi_k + \frac{1}{2}\Delta\phi_k^{\mathrm{T}}\, \nabla_{\phi_k}^2\mathcal{L}(\phi_k; y_k)\big|_{\phi_k=\phi_k^\tau} \Delta\phi_k + \mathcal{O}\left(\Delta{\phi_k}^3\right). \tag{30}$$

Then, it is relatively straightforward to obtain

$$\phi_k^* = \arg\max_{\phi_k} \mathcal{L}(\phi_k; y_k) = \arg\max_{\Delta\phi_k \to 0} \nabla_{\phi_k}\mathcal{L}(\phi_k; y_k) \cdot \Delta\phi_k + \frac{1}{2}\Delta\phi_k^{\mathrm{T}} \cdot \nabla_{\phi_k}^2\mathcal{L}(\phi_k; y_k) \cdot \Delta\phi_k. \tag{31}$$

The iterative formula for the variational parameter ϕ_k is then expressed as

$$\phi_k^{\tau+1} = \phi_k^\tau - \delta_\tau \cdot \left(\nabla_{\phi_k}^2\mathcal{L}(\phi_k; y_k)\right)^{-1} \nabla_{\phi_k}\mathcal{L}(\phi_k; y_k)\Big|_{\phi_k=\phi_k^\tau}, \tag{32}$$

where δ_τ is an adjustable iteration step.

So far, we still need to determine how to calculate gradients of the lower bound with respect to ϕ_k, which is slightly problematic. Ignoring the coupling effect of the variational parameters, we can decompose (32) into

$$\hat{x}_k^{\tau+1} = \hat{x}_k^{\tau} - \delta_\tau \cdot \left(\nabla^2_{\hat{x}_k} \mathcal{L}\left(\phi_k; y_k\right)\right)^{-1} \nabla_{\hat{x}_k} \mathcal{L}\left(\phi_k; y_k\right)\Big|_{\hat{x}_k=\hat{x}_k^{\tau}}, \tag{33}$$

and

$$P_k^{\tau+1} = P_k^{\tau} - \delta_\tau \cdot \left(\nabla^2_{P_k} \mathcal{L}\left(\phi_k; y_k\right)\right)^{-1} \nabla_{P_k} \mathcal{L}\left(\phi_k; y_k\right)\Big|_{P_k=P_k^{\tau}}, \tag{34}$$

thus improving calculation convenience. Subsequently, we individually address the computation of the variational parameters $\hat{x}_k$ and P_k.

Natural gradient estimator. In traditional variational Bayesian filter (15), the natural gradient estimator is used and it is assumed that the approximating posterior $q_{\phi_k^\tau}\left(x_k|y_{1:k}\right)$ at the τ-th iteration is the prior at the $(\tau+1)$-th iteration, which means

$$p\left(x_k|y_{1:k-1}\right) \approx q_{\phi_k^\tau}\left(x_k|y_{1:k}\right), \tag{35}$$

where ϕ_k^τ is the parameter of variational distribution at τ-th iteration. Noting that $\Delta\phi_k = \phi_k - \phi_k^\tau \to 0$, we obtain

$$\begin{cases} \nabla_{\phi_k} \mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|x_k\right)\right] = \nabla_{\phi_k} \mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|\hat{x}_k^\tau\right)\right] + \mathcal{O}\left(\Delta\phi_k\right), \\ \nabla_{\phi_k} \mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\left(x_k|y_{1:k}\right)\right) = \nabla^2_{\phi_k} \mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right) \cdot \Delta\phi_k + \mathcal{O}\left(\Delta{\phi_k}^2\right). \end{cases} \tag{36}$$

For the sake of brevity, denote the notation

$$G\left(\phi_k^\tau\right) = \nabla^2_{\phi_k} \mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right).$$

Then, according to (33) and (34), the iterative formula of the variational parameter $\{\hat{x}_k, P_k\}$ is

$$\begin{cases} \hat{x}_k^{\tau+1} = \hat{x}_k^{\tau} + G^{-1}\left(\hat{x}_k^{\tau}\right) \cdot \nabla_{\hat{x}_k} \mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|\hat{x}_k^{\tau}\right)\right], \\ P_k^{\tau+1} = P_k^{\tau} + G^{-1}\left(P_k^{\tau}\right) \cdot \nabla_{P_k} \mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|\hat{x}_k^{\tau}\right)\right]. \end{cases} \tag{37}$$

Assuming that the noises satisfy the Gaussian condition ($w_k \sim \mathcal{N}\left(0, Q_k\right), v_k \sim \mathcal{N}\left(0, R_k\right)$), via careful derivation (Appendix C), we easily obtain

$$\begin{cases} \nabla_{\hat{x}_k} \mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|\hat{x}_k^{\tau}\right)\right] \approx H^{\mathrm{T}}_{\hat{x}_k^{\tau}} R_k^{-1}\left(y_k - h_k\left(\hat{x}_k^{\tau}\right)\right), \\ \nabla_{P_k} \mathbb{E}_{q_{\phi_k}}\left[\log p\left(y_k|\hat{x}_k^{\tau}\right)\right] \approx -\dfrac{1}{2} H^{\mathrm{T}}_{\hat{x}_k^{\tau}} R_k^{-1} H_{\hat{x}_k^{\tau}}, \end{cases} \qquad H_{\hat{x}_k^{\tau}} = \left.\frac{\partial h_k\left(x_k\right)}{\partial x_k^{\mathrm{T}}}\right|_{x_k=\hat{x}_k^{\tau}}, \tag{38}$$

and

$$\begin{cases} G\left(\hat{x}_k^{\tau}\right) = \nabla^2_{\hat{x}_k} D_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right) = \left(P_k^{\tau}\right)^{-1}, \\ G\left(P_k^{\tau}\right) = \nabla^2_{P_k} D_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right) \approx \dfrac{1}{2}\left(P_k^{\tau}\right)^{-1} \otimes \left(P_k^{\tau}\right)^{-1}, \end{cases} \tag{39}$$

where $\otimes$ denotes the Kronecker product, producing the operator $\left(B^{\mathrm{T}} \otimes A\right) X = AXB$. Thus, Eq. (37) can be calculated in detail.

Preparation for Monte-Carlo gradient estimator. The natural gradient estimator heavily relies on the Gaussian assumption, resulting in an analytical gradient expression. However, KDE weakens the Gaussian assumption, thus hindering gradient calculation. Therefore, we introduce the Monte-Carlo gradient estimator to solve this problem. Before commencing, certain conclusions about the derivative of q_{ϕ_k} are essential, one of which is

$$\nabla_{\phi_k} q_{\phi_k} = q_{\phi_k} \cdot \nabla_{\phi_k} \log q_{\phi_k}. \tag{40}$$

Then, with (16),

$$\nabla_{\phi_k} \log q_{\phi_k} = -\frac{1}{2} \nabla_{\phi_k} \left(\log\left|P_k\right| + \left(x_k - \hat{x}_k\right)^{\mathrm{T}} P_k^{-1}\left(x_k - \hat{x}_k\right)\right), \tag{41}$$

which means

$$\begin{cases} \nabla_{\hat{x}_k} \log q_{\phi_k} = P_k^{-1}\left(x_k - \hat{x}_k\right), \\ \nabla_{P_k} \log q_{\phi_k} = -\dfrac{1}{2}\left(P_k^{-1} - P_k^{-1}\left(x_k - \hat{x}_k\right)\left(x_k - \hat{x}_k\right)^{\mathrm{T}} P_k^{-1}\right). \end{cases} \tag{42}$$

With (40)–(42), the gradient of $\mathcal{L}(\phi_k; y_k)$ can be analyzed thoroughly.

Monte-Carlo gradient estimator. According to (22),

$$\nabla_{P_k} p(y_k|x_k) = 0, \quad \nabla_{P_k} p(x_k|x_{k-1}) = 0, \quad \nabla_{P_k} p(x_k|y_{1:k-1}) = 0. \tag{43}$$

Therefore, we only need to consider the general gradient estimator of $\mathbb{E}_{q_{\phi_k}}[g(x_k)]$ and $\mathbb{D}_{KL}(q_{\phi_k}|g(x_k))$. Combining (40), we obtain

$$\begin{cases} \nabla_{\phi_k} \mathbb{E}_{q_{\phi_k}}[g(x_k)] = \mathbb{E}_{q_{\phi_k}}[g \cdot (\nabla_{\phi_k} \log q_{\phi_k})], \\ \nabla^2_{\phi_k} \mathbb{E}_{q_{\phi_k}}[g(x_k)] = \mathbb{E}_{q_{\phi_k}}\left[g \cdot \left((\nabla_{\phi_k} \log q_{\phi_k})^2 + \nabla^2_{\phi_k} \log q_{\phi_k}\right)\right], \\ \nabla_{\phi_k} \mathbb{D}_{KL}(q_{\phi_k}|g(x_k)) = \mathbb{E}_{q_{\phi_k}}\left[\left(\log \dfrac{q_{\phi_k}}{g} + 1\right) \cdot \nabla_{\phi_k} \log q_{\phi_k}\right], \\ \nabla^2_{\phi_k} \mathbb{D}_{KL}(q_{\phi_k}|g(x_k)) = \mathbb{E}_{q_{\phi_k}}\left[\left((\nabla_{\phi_k} \log q_{\phi_k})^2 + \nabla^2_{\phi_k} \log q_{\phi_k}\right) \cdot \left(\log \dfrac{q_{\phi_k}}{g} + 1\right) + (\nabla_{\phi_k} \log q_{\phi_k})^2\right]. \end{cases} \tag{44}$$

According to the Monte-Carlo method, the approximate calculation formula for the expectation can be represented as

$$\mathbb{E}_{q_{\phi_k}}[\gamma(\phi_k)] \approx \frac{1}{S} \sum_{s=1}^{S} \gamma\left(\xi^{(s)}\right), \quad \xi^{(s)} \sim q_{\phi_k}(x_k|y_{1:k}), \tag{45}$$

where $\xi^{(s)}$ is a series of generated samples that follow q_{ϕ_k}. Then, combining (44) and (45), we calculate the required gradient (using the reparameterization technique in Appendix D). Remarkably, the gradient estimator of $\nabla_{\phi_k} \mathbb{E}_{q_{\phi_k}}[g(x_k)]$ is

$$\begin{cases} \mathbb{E}_{q_{\phi_k}}[g \cdot \nabla_{\hat{x}_k} \log q_{\phi_k}] = \dfrac{1}{S} \sum_{s=1}^{S} g \cdot \nabla_{\hat{x}_k} \log q_{\phi_k} = -\dfrac{1}{S} \sum_{s=1}^{S} g \cdot \zeta^{(s)}, \\ \mathbb{E}_{q_{\phi_k}}[g \cdot \nabla_{P_k} \log q_{\phi_k}] = -\dfrac{1}{2S} \sum_{s=1}^{S} g \cdot P_k^{-1} \cdot \left(I - \zeta^{(s)} \left(\zeta^{(s)}\right)^{\mathrm{T}} P_k^{-1}\right), \end{cases} \quad \zeta^{(s)} \sim \mathcal{N}(0, I). \tag{46}$$

Gradient of adaptive weight function. According to (44), given the adaptive weight function, we only need to consider the gradient of $\Psi(x_k)$. Through calculation, we obtain

$$\begin{cases} \varphi'(z) = \dfrac{d}{dz}\varphi(z) = \alpha^* \cdot \mathrm{e}^{-k(1-c/|z|)^2} \cdot \left(1 - \dfrac{c}{|z|}\right) \cdot \dfrac{\mathrm{sgn}(z)}{z^2} \cdot \mathbb{I}(|z| > c), \\ \varphi''(z) = \dfrac{d^2}{dz^2}\varphi(z) = \alpha^* \cdot \mathrm{e}^{-k(1-c/|x|)^2} \cdot \dfrac{1}{z^3} \cdot \left(-2\dfrac{kc}{|z|}\left(1 - \dfrac{c}{|z|}\right)^2 - 2 + 3\dfrac{c}{|z|}\right) \cdot \mathbb{I}(|z| > c), \end{cases} \tag{47}$$

where $\alpha^* = -2kc/\left(1 - \mathrm{e}^{-k}\right)$. Because $\Psi(x_k)$ and P_k are independent, we only need to consider their gradient with respect to x_k, which mean

$$\begin{aligned} \nabla_{x_k}\left(\varphi\left(r_k^{(i),2}\right) \cdot r_k^{(i),2}\right) &= \nabla_{x_k}\varphi\left(r_k^{(i),2}\right) \cdot r_k^{(i),2} + \varphi^{(i)} \cdot \nabla_{x_k} r_k^{(i),2} \\ &= -2 \cdot \hat{R}_k^{-1/2} \cdot r_k^{(i)} \cdot \left(\varphi'\left(r_k^{(i),2}\right) \cdot r_k^{(i),2} + \varphi\left(r_k^{(i),2}\right)\right) \cdot \nabla_{x_k} h_k^{(i)}, \end{aligned} \tag{48}$$

and

$$\begin{aligned} &\nabla^2_{x_k}\left(\varphi\left(r_k^{(i),2}\right) \cdot r_k^{(i),2}\right) \\ &= 2\hat{R}_k^{-1} \left(\begin{array}{r} \left(\nabla_{x_k} h_k^{(i)} \cdot \nabla^{\mathrm{T}}_{x_k} h_k^{(i)}\right)\left(\varphi\left(r_k^{(i),2}\right) + 5r_k^{(i),2}\varphi'\left(r_k^{(i),2}\right) + 2\left(r_k^{(i),2}\right)^2 \varphi''\left(r_k^{(i),2}\right)\right) \\ -\hat{R}_k^{1/2}\left(r_k^{(i)} \cdot \left(\varphi'\left(r_k^{(i),2}\right) \cdot r_k^{(i),2} + \varphi\left(r_k^{(i),2}\right)\right) \cdot \nabla^2_{x_k} h_k^{(i)}\right) \end{array} \right). \end{aligned} \tag{49}$$

Thus, $\nabla_{x_k}\Psi(x_k)$ and $\nabla^2_{x_k}\Psi(x_k)$ can be calculated according to (48) and (49).

Overall. So far, all essential conditions for conducting AW-VBF calculations have been provided. Its process is summarized in Algorithm 1.

Algorithm 1 AW-VBF algorithm with Monte-Carlo gradient estimator.

Initialization: State estimation $\hat{x}_0$, estimation error covariance P_0, termination threshold ε, parameters L, λ, c, k, and iteration N.
for all $k = 1, 2, \ldots, K$ **do**
  Compute initial state estimation: $\hat{x}_k^0 = f_{k|k-1}(\hat{x}_{k-1})$;
  Compute initial covariance matrix:
  $P_k^0 = F_{k|k-1} P_{k-1} F_{k|k-1}^{\mathrm{T}} + Q_{k-1}, F_{k|k-1} = \partial f_{k|k-1}(x)/\partial x^{\mathrm{T}}\Big|_{x=\hat{x}_k^0}$;
  for all $\tau = 1, 2, \ldots, N$ and $e > \varepsilon$ **do**
    Generate random samples: $\zeta^{(s)} \sim \mathcal{N}(0, I)$;
    Compute the gradient of $\mathbb{D}_{KL}\left(q_{\phi_k} | p(x_k|y_{1:k-1})\right)$ at ϕ_k^τ according to (44) and (46);
    Compute the gradient of $\mathbb{D}_M(\phi_k; y_k)$ at ϕ_k^τ according to (48) and (49);
    Update the variational parameter $\phi_k^{\tau+1}$ according to (32);
    Compute relevant error $e = \left\| \left(\hat{x}_k^{\tau+1} - \hat{x}_k^\tau\right) / \hat{x}_k^\tau \right\|$;
  end for
Output: $\hat{x}_k = \hat{x}_k^\tau$ and $P_k = P_k^\tau$;
  Compute prediction residuals: $\hat{x}_k - f_{k|k-1}(\hat{x}_{k-1})$;
  Update the KDE of distribution $p(x_{k+1}|x_k)$ to $\frac{1}{L}\sum_{i=k-L+1}^{k} \kappa\left(x_k - \mu_x^{(i)}; \lambda \cdot I\right)$;
end for

4 Simulation

4.1 Preparation

The simulation was designed for a target-tracking [4] scenario. The dynamics of the moving target was described using a constant-velocity model as

$$\begin{bmatrix} p_k \\ \dot{p}_k \end{bmatrix} = \begin{bmatrix} I & d_t \cdot I \\ 0 & I \end{bmatrix} \begin{bmatrix} p_{k-1} \\ \dot{p}_{k-1} \end{bmatrix} + w_{k-1}, \tag{50}$$

where $d_t = t_k - t_{k-1}$ is the discretization time interval, w_{k-1} is a process noise with appropriate dimensions, $p_k = [x_k, y_k, z_k]^{\mathrm{T}}$ and $\dot{p}_k = [\dot{x}_k, \dot{y}_k, \dot{z}_k]^{\mathrm{T}}$ is the position and velocity of the target, respectively, at time t_k.

The measurement primarily consisted of the radial distance (R) and radial velocity ($\dot{R}$) between the source and the target. Accordingly, the measurement function was

$$\begin{cases} R(p_k, s_i) = \sqrt{(x_k^p - x_i^s)^2 + (y_k^p - y_i^s)^2 + (z_k^p - z_i^s)^2} + v_{k,i}^{(1)}, \\ \dot{R}(p_k, s_i) = \left(\dot{x}_k^p (x_k^p - x_i^s) + \dot{y}_k^p (y_k^p - y_i^s) + \dot{z}_k^p (z_k^p - z_i^s)\right) / R(p_k, s_i) + v_{k,i}^{(2)}, \end{cases} \tag{51}$$

where $s_i = [x_i^s, y_i^s, z_i^s]^{\mathrm{T}}$ represents the position of the i-th measurement source, $v_{k,i}^{(1)}$ and $v_{k,i}^{(2)}$ are the noise of the corresponding measurement, at time t_k.

4.2 Parameters and visualization

In the simulation, six range-velocity radars were deployed for obtaining the measurements of R and $\dot{R}$, and the specific positional parameters are detailed in Table 1. Then, noises and anomalies with distinct characteristics were simulated to test the performance of diverse filters.

(1) Regarding the process noise w_k, Gaussian noise and non-Gaussian noise are separately simulated, where Gaussian noise is designated as $w_k \sim \mathcal{N}(w_k; 0, \Lambda_1)$ and non-Gaussian noise is designated as $w_k \sim \frac{1}{2}(\mathcal{N}(w_k; \mu_1, \Lambda_1) + \mathcal{N}(w_k; -\mu_1, \Lambda_1))$.

(2) For normal measurements, its noise follows a normal distribution, that is, $v_k \sim \mathcal{N}(v_k; 0, \Lambda_2)$. For anomalies, at most one measurement is abnormal at each moment, and its occurrence probability is set to $p = 0.2$. Moreover, the range anomaly had an amplitude of 75 and a standard deviation of 15, whereas the velocity anomaly had an amplitude of 7.5 and a standard deviation of 1.5.

In particular, the relevant parameters related to noise and anomalies are as follows:

$$\begin{cases} \mu_1 = [0.5, 0.5, 0.5, 0, 0, 0]^{\mathrm{T}}, v_k = \left[v_{k,1}^{(1)}, \ldots, v_{k,6}^{(1)}, v_{k,1}^{(2)}, \ldots, v_{k,6}^{(2)}\right]^{\mathrm{T}}, \\ \Lambda_1 = \mathrm{diag}\left(\left[10^{-2}, 10^{-2}, 10^{-4}, 10^{-2}, 10^{-2}, 2.5 \times 10^{-5}\right]\right), \\ \Lambda_2 = \mathrm{diag}\left(\left[10^3, 10^3, 10^3, 10^3, 10^3, 10^3, 10^{-2}, 10^{-2}, 10^{-2}, 10^{-2}, 10^{-2}, 10^{-2}\right]\right). \end{cases}$$

Furthermore, the simulation scene and some data details are visualized here. The target trajectory and measurement source deployment are shown in Figure 3, and the change curves of the velocity components (with Gaussian

Table 1 Parameter settings related to simulation scenarios.

Parameter	Value	Unit	Parameter	Value	Unit
p_0	$(125, -3000, 100)^{\mathrm{T}}$	m	d_t	0.2	s
$\dot{p}_0$	$(0, 60, -0.75)^{\mathrm{T}}$	$\mathrm{m}\cdot\mathrm{s}^{-1}$	a	100	–
s_1–s_6	$(\pm a, \pm b, 0)^{\mathrm{T}}, (\pm a, 0, 0)^{\mathrm{T}}$	m	b	2000	–

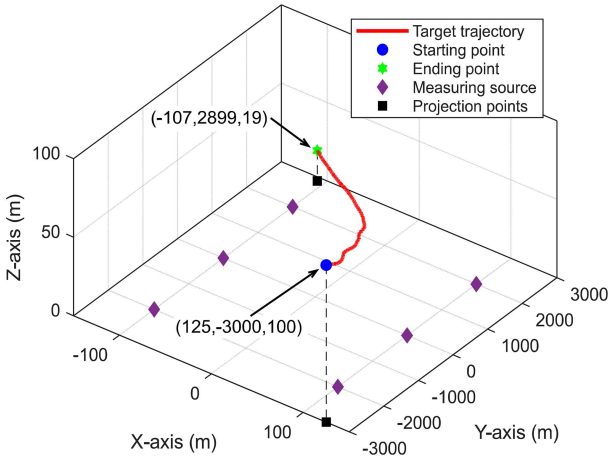

Figure 3 (Color online) Trajectory and measurement source position.

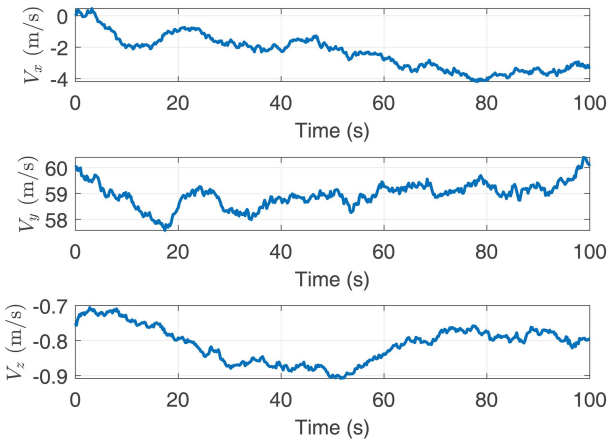

Figure 4 (Color online) Change curves of velocity components.

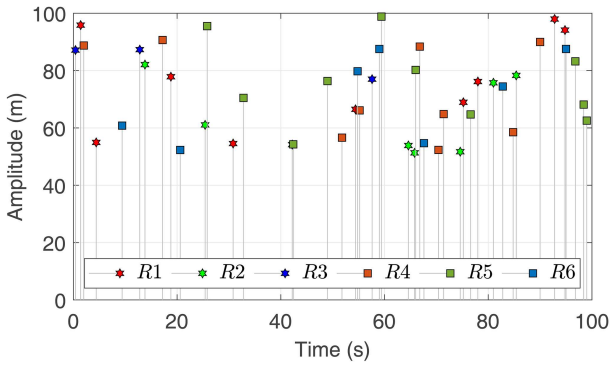

Figure 5 (Color online) Anomalies injected into radial distance R.

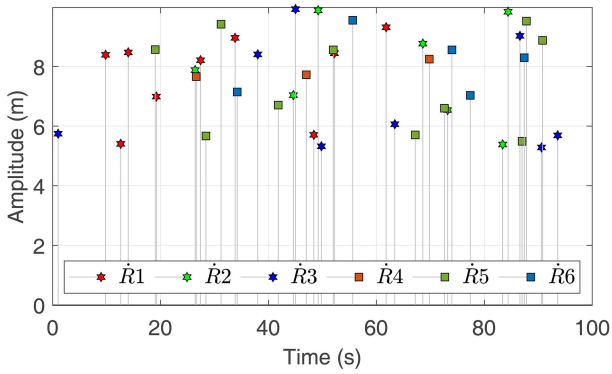

Figure 6 (Color online) Anomalies injected into radial distance $\dot{R}$.

process noise and normal measurements) are shown in Figure 4. In addition, Figures 5 and 6 present the anomalies in radial distance and radial distance measurements, respectively, which are injected in the simulation.

4.3 Performance

We compare the performance of the AW-VBF algorithm with that of three related algorithms: EKF, UKF, and PF. The EKF, a widely used linearization method, is a popular traditional baseline. The UKF can effectively handle nonlinear problems by using the unscented transformation instead of linearization. The PF describes the distribution of state variables through nonlinear systems through a large number of random samples; it is a representative method for dealing with nonlinear and non-Gaussian problems.

The residual between the estimated state and the true state, directly reflecting the performance of the filter, was utilized as an evaluation metric. Position and velocity were evaluated separately to ensure dimensional consistency, and the relevant metrics (at time t_k) were

$$\mathrm{RMSE}\left(p_k\right) = \left\|p_k - \hat{p}_k\right\|_2^2, \quad \mathrm{RMSE}\left(\dot{p}_k\right) = \left\|\dot{p}_k - \hat{\dot{p}}_k\right\|_2^2.$$

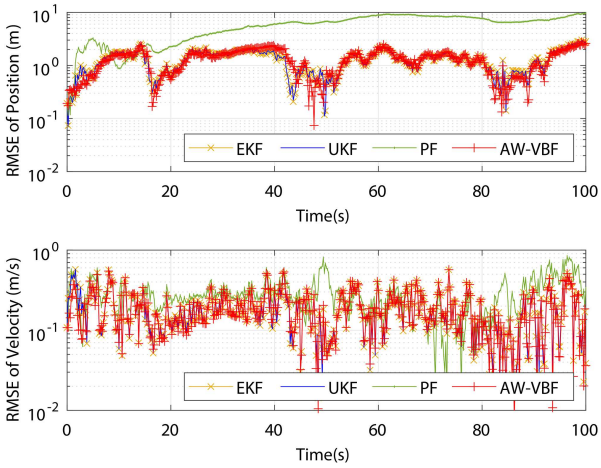

Figure 7 (Color online) Case 1: system state estimation results when the process noise is Gaussian and no measurement anomalies exist.

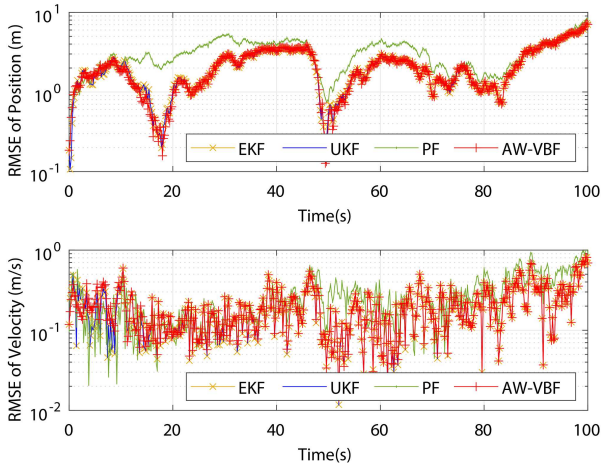

Figure 8 (Color online) Case 2: system state estimation results when the process noise is non-Gaussian and no measurement anomalies exist.

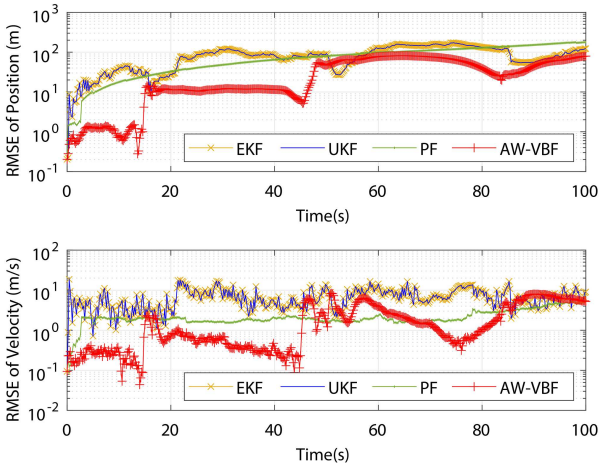

Figure 9 (Color online) Case 3: system state estimation results when the process noise is Gaussian and anomalies exist in the measurements.

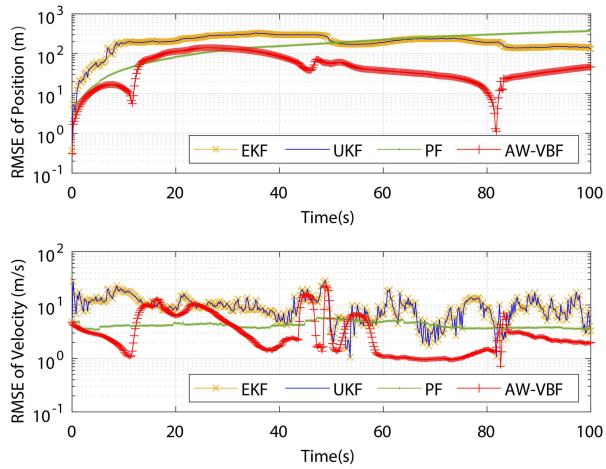

Figure 10 (Color online) Case 4: system state estimation results when the process noise is non-Gaussian and anomalies exist in the measurements.

In the simulation, the parameters of AW-VBF are selected as follows:

$$L = 100, \quad \lambda = 0.03, \quad k = 10, \quad c = 3 \times \sqrt{10^3}.$$

The simulation results of our proposed method (AW-VBF) and the compared methods are in Figures 7–10. As shown in Figure 7, the AW-VBF maintains consistency with the classical filters in terms of Gaussian noise and normal measurements. Figure 8 illustrates the effect of the non-Gaussian process noise on the filters, showing acceptable degradation. Nevertheless, our method is less affected by this factor because of the precise estimation of the process noise. Figure 9 illustrates filter performance in the presence of measurement anomalies. Because the classical filters do not handle measurement anomalies, their performance degrades severely, resulting in a substantial amplification (approximately 65 times) of position estimation errors and an amplification (approximately 35 times) of velocity estimation errors.

Tables 2 and 3 present statistical data regarding the position and velocity estimation residuals, respectively. By contrast, our proposed method (AW-VBF) effectively alleviates this decline, improving position estimation accuracy by 64.9% to 66.6% and velocity estimation accuracy by 25.6% to 76.8% relative to the classical methods. The effectiveness of our approach is further validated in Figure 10, where position estimation accuracy improves by at least 61.1% and velocity estimation accuracy by at least 5.47%.

Table 2 Position estimation performance (Mean ± Std, m) of AW-VBF and competing methods.

Cases	AW-VBF	EKF	UKF	PF
Case 1	1.285 ± 0.611	1.296 ± 0.581	1.292 ± 0.579	4.814 ± 1.794
Case 2	2.017 ± 1.173	3.251 ± 1.838	3.560 ± 1.927	7.436 ± 6.281
Case 3	28.248 ± 17.824	84.566 ± 44.902	84.564 ± 44.899	80.534 ± 45.976
Case 4	51.934 ± 31.461	206.414 ± 71.854	206.412 ± 71.856	133.564 ± 76.713

Table 3 Velocity estimation performance (Mean ± Std, m/s) of AW-VBF and competing methods.

Cases	AW-VBF	EKF	UKF	PF
Case 1	0.185 ± 0.106	0.190 ± 0.110	0.189 ± 0.110	0.284 ± 0.162
Case 2	0.251 ± 0.147	0.371 ± 0.201	0.378 ± 0.210	0.720 ± 0.403
Case 3	1.579 ± 1.087	6.795 ± 3.456	6.794 ± 3.455	2.122 ± 0.837
Case 4	3.110 ± 2.454	9.494 ± 4.534	9.494 ± 4.534	3.291 ± 1.111

5 Conclusion

We propose the AW-VBF to generate an approximate variational distribution for the true state posterior distribution. In the AW-VBF, the kernel density function is used to estimate unknown and non-Gaussian system noise, and a novel weighted loss function with excellent properties replaces the likelihood function, thereby achieving robust estimation. Thus, the AW-VBF can maintain its performance in the presence of non-Gaussian noise and measurement anomalies. Simulation and analysis results show that the proposed method outperforms existing methods (EKF, UKF, and PF).

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62203458) and the Postgraduate Scientific Research Innovation Project of Hunan Province (Grant Nos. CX20220024, CX20230014).

References

1 Elorza Casas C A, Valipour M, Ricardez Sandoval L A. Multi-scenario and multi-stage robust NMPC with state estimation application on the Tennessee-Eastman process. Control Eng Pract, 2023, 139: 105635
2 Pourtakdoust S H, Mehrjardi M F, Hajkarim M H. Attitude estimation and control based on modified unscented Kalman filter for gyro-less satellite with faulty sensors. Acta Astronaut, 2022, 191: 134–147
3 Zhao L, Liu Z, Ji G. Lithium-ion battery state of charge estimation with model parameters adaptation using H_∞ extended Kalman filter. Control Eng Pract, 2018, 81: 114–128
4 Ristic B, Houssineau J, Arulampalam S. Target tracking in the framework of possibility theory: The possibilistic Bernoulli filter. Inf Fusion, 2020, 62: 81–88
5 Li T, Liang H, Xiao B, et al. Finite mixture modeling in time series: A survey of Bayesian filters and fusion approaches. Inf Fusion, 2023, 98: 101827
6 Bai W Y, Xue W C, Huang Y, et al. On extended state based Kalman filter design for a class of nonlinear time-varying uncertain systems. Sci China Inf Sci, 2018, 61: 042201
7 Bai Y, Yan B, Zhou C, et al. State of art on state estimation: Kalman filter driven by machine learning. Annu Rev Control, 2023, 56: 100909
8 Xiong K, Zhang H Y, Chan C W. Performance evaluation of UKF-based nonlinear filtering. Automatica, 2006, 42: 261–270
9 Arasaratnam I, Haykin S. Cubature Kalman filters. IEEE Trans Automat Contr, 2009, 54: 1254–1269
10 Jia B, Xin M, Cheng Y. High-degree cubature Kalman filter. Automatica, 2013, 49: 510–518
11 Alspach D L. Gaussian sum approximations in nonlinear filtering and control. Inf Sci, 1974, 7: 271–290
12 Steckenrider J J, Furukawa T. Simultaneous estimation and modeling of nonlinear, non-Gaussian state-space systems. Inf Sci, 2021, 578: 621–643
13 Feng X, Wu C, Ge Q. Sequential fusion filtering based on minimum error entropy criterion. Inf Fusion, 2024, 104: 102193
14 Liu Y, Wang H, Hou C. UKF based nonlinear filtering using minimum entropy criterion. IEEE Trans Signal Process, 2013, 61: 4988–4999
15 Liu W, Pokharel P P, Principe J C. Correntropy: Properties and applications in non-Gaussian signal processing. IEEE Trans Signal Process, 2007, 55: 5286–5298
16 Li X L, Adali T. Complex-valued linear and widely linear filtering using MSE and Gaussian entropy. IEEE Trans Signal Process, 2012, 60: 5672–5684
17 Chen B, Xing L, Zhao H, et al. Generalized correntropy for robust adaptive filtering. IEEE Trans Signal Process, 2016, 64: 3376–3387
18 Zhang X, Dees B S, Li C, et al. Analysis of least stochastic entropy adaptive filters for noncircular Gaussian signals. IEEE Trans Circ Syst II Express Briefs, 2020, 67: 1364–1368
19 Yang B, Du B, Li N, et al. Centered error entropy-based variational Bayesian adaptive and robust Kalman filter. IEEE Trans Circ Syst II Express Briefs, 2022, 69: 5179–5183
20 Tian B, Wang Y, Guo L. Entropy optimization based filtering for non-Gaussian stochastic systems. Sci China Inf Sci, 2017, 60: 120203
21 Li Z, Xing L, Chen B. Adaptive filtering with quantized minimum error entropy criterion. Signal Process, 2020, 172: 107534
22 Feng Z, Wang G, Peng B, et al. Distributed minimum error entropy Kalman filter. Inf Fusion, 2023, 91: 556–565
23 Qian G, Liu J, Qiu C, et al. Minimum total complex error entropy for adaptive filter. Expert Syst Appl, 2024, 237: 121522
24 Lin L K, Xu H, Sheng W D, et al. Multi-target state-estimation technique for the particle probability hypothesis density filter. Sci China Inf Sci, 2012, 55: 2318–2328
25 Doucet A, Godsill S, Andrieu C. On sequential Monte Carlo sampling methods for Bayesian filtering. Stat Comput, 2000, 10: 197–208
26 Zhang G Y, Cheng Y M, Yang F, et al. Design of an adaptive particle filter based on variance reduction technique. Acta Automat Sin, 2010, 36: 1020–1024
27 Wang H, Li H, Fang J, et al. Robust Gaussian Kalman filter with outlier detection. IEEE Signal Process Lett, 2018, 25: 1236–1240
28 Zou L, Wang Z, Geng H, et al. Set-membership filtering subject to impulsive measurement outliers: A recursive algorithm. IEEE CAA J Autom Sin, 2021, 8: 377–388

29 Yang S, Fu H. Variational robust filter for a class of stochastic systems with false and missing measurements. J Franklin Inst, 2024, 361: 106941
30 Zhong S, Wang Z, Wang G, et al. Robust adaptive filtering based on M-estimation-based minimum error entropy criterion. Inf Sci, 2024, 658: 120026
31 Wang W, Tse C K, Wang S. Generalized correntropy sparse Gauss-Hermite quadrature filter for epidemic tracking on complex networks. IEEE Trans Syst Man Cyber Syst, 2022, 52: 2770–2778
32 Fang H, Haile M A, Wang Y. Robust extended Kalman filtering for systems with measurement outliers. IEEE Trans Contr Syst Technol, 2022, 30: 795–802
33 Sarkka S, Nummenmaa A. Recursive noise adaptive Kalman filtering by variational Bayesian approximations. IEEE Trans Automat Contr, 2009, 54: 596–600
34 Li K, Chang L, Hu B. A variational Bayesian-based unscented Kalman filter with both adaptivity and robustness. IEEE Sens J, 2016, 16: 6966–6976
35 Yu X, Li J, Xu J. Nonlinear filtering in unknown measurement noise and target tracking system by variational Bayesian inference. Aerospace Sci Tech, 2019, 84: 37–55
36 Ma C, Pan S, Gao W, et al. Variational Bayesian-based robust adaptive filtering for GNSS/INS tightly coupled positioning in urban environments. Measurement, 2023, 223: 113668
37 Dong P, Jing Z, Leung H, et al. Variational Bayesian adaptive cubature information filter based on Wishart distribution. IEEE Trans Automat Contr, 2017, 62: 6051–6057
38 Huang Y, Zhang Y, Shi P, et al. Variational adaptive Kalman filter with Gaussian-Inverse-Wishart mixture distribution. IEEE Trans Automat Contr, 2021, 66: 1786–1793
39 Yang B, Wang H, Shi Z. Variational Bayesian and generalized maximum-likelihood based adaptive robust nonlinear filtering framework. Signal Process, 2024, 215: 109271
40 Fox C W, Roberts S J. A tutorial on variational Bayesian inference. Artif Intell Rev, 2012, 38: 85–95
41 Lai Y, Guan W, Luo L, et al. Bayesian estimation of inverted beta mixture models with extended stochastic variational inference for positive vector classification. IEEE Trans Neural Netw Learn Syst, 2024, 35: 6948–6962
42 Hu Y, Wang X, Pan Q, et al. Variational Bayesian Kalman filter using natural gradient. Chin J Aeronaut, 2022, 35: 1–10
43 Liu Y, Liu J, Xu C A, et al. Fully distributed variational Bayesian non-linear filter with unknown measurement noise in sensor networks. Sci China Inf Sci, 2020, 63: 210202
44 Huang W, Fu H, Zhang W. A novel robust variational Bayesian filter for unknown time-varying input and inaccurate noise statistics. IEEE Sensors Letters, 2023, 7: 1–4
45 Van Aelst S, Willems G, Zamar R H. Robust and efficient estimation of the residual scale in linear regression. J Multivariate Anal, 2013, 116: 278–296
46 Zou Y, Chan S C, Ng T S. Least mean m-estimate algorithms for robust adaptive filtering in impulse noise. IEEE Trans Circ Syst II Analog Digit Signal Process, 2000, 47: 1564–1569
47 Petrus P. Robust Huber adaptive filter. IEEE Trans Signal Process, 1999, 47: 1129–1133
48 Wei J H, He Z M, Wang J Q, et al. Fault detection based on multi-dimensional KDE and Jensen-Shannon divergence. Entropy, 2021, 23: 266
49 Zuo Y, Zuo H. Weighted least squares regression with the best robustness and high computability. Axioms, 2024, 13: 295

Appendix A Specific example for Bayesian filter with linear Gaussian assumption

We consider the discrete linear Gaussian system

$$\begin{cases} x_k = Fx_{k-1} + w_{k-1}, \\ y_k = Hx_k + v_k, \end{cases} \quad w_k \sim \mathcal{N}(0, Q_k)\,,\ v_k \sim \mathcal{N}(0, R_k)\,, \tag{A1}$$

where F and H are the state transition matrix and state measurement matrix, respectively, with appropriate dimensions. We assume that the state at time t_{k-1} follows a normal distribution with a mean of $\hat{x}_{k-1}$ and a covariance of P_{k-1}, which means

$$p(x_{k-1}|y_{1:k-1}) = \mathcal{N}(x_{k-1}; \hat{x}_{k-1}, P_{k-1})\,. \tag{A2}$$

Therefore,

$$\begin{cases} p(x_k|y_{1:k}) = \mathcal{N}\left(x_k; F\hat{x}_{k-1} + K_k(y_k - HF\hat{x}_{k-1}),\ P_{k|k-1} - K_k\left(HP_{k|k-1}H^{\mathrm{T}} + R_k\right)K_k^{\mathrm{T}}\right), \\ K_k = P_{k|k-1}H^{\mathrm{T}}\left(HP_{k|k-1}H^{\mathrm{T}} + R_k\right)^{-1},\ P_{k|k-1} = FP_{k-1}F^{\mathrm{T}} + Q_{k-1}. \end{cases} \tag{A3}$$

Simplifying (A3) can yield the classic KF, which is

$$\begin{cases} \hat{x}_k = F\hat{x}_{k-1} + K_k(y_k - HF\hat{x}_{k-1}), \\ P_k = P_{k|k-1} - K_k\left(HP_{k|k-1}H^{\mathrm{T}} + R_k\right)K_k^{\mathrm{T}}. \end{cases} \tag{A4}$$

Appendix B Computation details for VBF with linear Gaussian assumption

With the assumption (A1), there is

$$\begin{aligned} &\mathbb{E}_{q_{\phi_k}}\left[\log p(y_k|x_k)\right] = \int \mathcal{N}(x_k; \hat{x}_k, P_k)\cdot\log\left[\mathcal{N}(y_k; Hx_k, R_k)\right]dx_k \\ &= -\frac{1}{2}\left(n_y\log 2\pi + \log|R_k| + \int \mathcal{N}(x_k; \hat{x}_k, P_k)\cdot(y_k - Hx_k)^{\mathrm{T}}R_k^{-1}(y_k - Hx_k)\,dx_k\right) \\ &= -\frac{1}{2}\left(n_y\log 2\pi + \log|R_k| + \int \mathcal{N}(x_k; 0, P_k)\cdot((y_k - H\hat{x}_k) - Hx_k)^{\mathrm{T}}R_k^{-1}((y_k - H\hat{x}_k) - Hx_k)\,dx_k\right). \end{aligned} \tag{B1}$$

Owing to the evidence $\int \mathcal{N}(x_k;0,P_k)\cdot \tilde{H}x_k dx_k=0$ and

$$\begin{aligned}&\int \mathcal{N}(x_k;0,P_k)\cdot x_k^{\mathrm{T}}H^{\mathrm{T}}R_k^{-1}Hx_k dx_k=\mathbb{E}_{\mathcal{N}(x_k;0,I)}\left[\left(R_k^{-1/2}HP_k^{1/2}x_k\right)^{\mathrm{T}}\left(R_k^{-1/2}HP_k^{1/2}x_k\right)\right]\\&=\mathrm{Tr}\left(\left(R_k^{-1/2}HP_k^{1/2}\right)^{\mathrm{T}}R_k^{-1/2}HP_k^{1/2}\right)=\mathrm{Tr}\left(R_k^{-1}HP_kH^{\mathrm{T}}\right).\end{aligned}\tag{B2}$$

Eq. (B1) can be rewritten as follows:

$$\mathbb{E}_{q_{\phi_k}}\left[\log p(y_k|x_k)\right]=-\frac{1}{2}\left(n_y\log 2\pi+\log|R_k|+(y_k-H\hat{x}_k)^{\mathrm{T}}R_k^{-1}(y_k-H\hat{x}_k)+\mathrm{Tr}\left(R_k^{-1}HP_kH^{\mathrm{T}}\right)\right).\tag{B3}$$

Similarly, it is easy to verify $p(x_k|y_{1:k-1})=\mathcal{N}(x_k;F\hat{x}_{k-1},P_{k|k-1})$ and

$$\begin{aligned}&\mathbb{D}_{KL}\left(q_{\phi_k}\middle|p(x_k|y_{1:k-1})\right)=\mathbb{E}_{q_{\phi_k}}\left[\log\frac{p(x_k|y_{1:k-1})}{q_{\phi_k}(x_k|y_{1:k})}\right]=\mathbb{E}_{q_{\phi_k}}\left[\log\frac{\mathcal{N}(x_k;F\hat{x}_{k-1},P_{k|k-1})}{\mathcal{N}(x_k;\hat{x}_k,P_k)}\right]\\&=-\frac{1}{2}\log\left|P_{k|k-1}\right|+\frac{1}{2}\log|P_k|+\mathbb{E}_{q_{\phi_k}}\left[(x_k-F\hat{x}_{k-1})^{\mathrm{T}}P_{k|k-1}^{-1}(x_k-F\hat{x}_{k-1})-(x_k-\hat{x}_k)^{\mathrm{T}}P_k^{-1}(x_k-\hat{x}_k)\right]\\&=-\frac{1}{2}\left(\log\left|P_{k|k-1}\right|-\log|P_k|+(x_k-F\hat{x}_{k-1})^{\mathrm{T}}P_{k|k-1}^{-1}(x_k-F\hat{x}_{k-1})+\mathrm{Tr}\left(P_{k|k-1}^{-1}P_k\right)-\mathrm{Tr}(I_{n_x\times n_x})\right).\end{aligned}\tag{B4}$$

Therefore, Eqs. (B3) and (B4) yield (17).

Appendix C Computation details for natural gradient estimator

According to the assumption (35), the $\mathbb{D}_{KL}\left(q_{\phi_k}\middle|p(x_k|y_{1:k-1})\right)$ is given as

$$\mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right)=\int q_{\phi_k}\cdot\log\frac{q_{\phi_k}}{q_{\phi_k^\tau}}dx_k=\int q_{\phi_k}\cdot\log q_{\phi_k}dx_k-\int q_{\phi_k}\cdot\log q_{\phi_k^\tau}dx_k.\tag{C1}$$

With the Taylor's series, $\log q_{\phi_k}$ can be expanded into

$$\log q_{\phi_k}=\log q_{\phi_k^\tau}+\left(\nabla_{\phi_k}\log q_{\phi_k}\right)^{\mathrm{T}}\Big|_{\phi_k=\phi_k^\tau}\Delta\phi_k+\frac{1}{2}\Delta\phi_k^{\mathrm{T}}\left(\nabla_{\phi_k}^2\log q_{\phi_k}\right)^{\mathrm{T}}\Big|_{\phi_k=\phi_k^\tau}\Delta\phi_k+O\left(\Delta\phi_k^3\right).\tag{C2}$$

At point ϕ_k^τ, so that Eq. (C1) can be rewritten as

$$\mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right)=\int q_{\phi_k}\cdot\left(\nabla_{\phi_k}\log q_{\phi_k}\right)^{\mathrm{T}}\Big|_{\phi_k=\phi_k^\tau}\Delta\phi_k dx_k+\frac{1}{2}\int q_{\phi_k}\cdot\Delta\phi_k^{\mathrm{T}}\left(\nabla_{\phi_k}^2\log q_{\phi_k}\right)^{\mathrm{T}}\Big|_{\phi_k=\phi_k^\tau}\Delta\phi_k dx_k+O\left(\Delta\phi_k^3\right).\tag{C3}$$

The first RHS term in (C3) becomes

$$\int q_{\phi_k}\cdot\left(\nabla_{\phi_k}\log q_{\phi_k}\right)^{\mathrm{T}}\Big|_{\phi_k=\phi_k^\tau}\Delta\phi_k dx_k=\int\left(\nabla_{\phi_k}q_{\phi_k}\right)^{\mathrm{T}}\Big|_{\phi_k=\phi_k^\tau}\Delta\phi_k dx_k=0.\tag{C4}$$

The second RHS term in (C3) can be reformulated as

$$\begin{aligned}&\frac{1}{2}\int q_{\phi_k}\cdot\Delta\phi_k^{\mathrm{T}}\left(\nabla_{\phi_k}^2\log q_{\phi_k}\right)\Big|_{\phi_k=\phi_k^\tau}\Delta\phi_k dx_k=\frac{1}{2}\Delta\phi_k^{\mathrm{T}}\int q_{\phi_k}\cdot\left(\nabla_{\phi_k}^2\log q_{\phi_k}\right)\Big|_{\phi_k=\phi_k^\tau}dx_k\Delta\phi_k\\&=\frac{1}{2}\Delta\phi_k^{\mathrm{T}}\int q_{\phi_k}\cdot\left(\frac{1}{q_{\phi_k}}\left(\nabla_{\phi_k}^2q_{\phi_k}\right)\Big|_{\phi_k=\phi_k^\tau}+\left(\nabla_{\phi_k}q_{\phi_k}\right)\left(\nabla_{\phi_k}q_{\phi_k}\right)^{\mathrm{T}}\Big|_{\phi_k=\phi_k^\tau}\right)dx_k\Delta\phi_k.\end{aligned}\tag{C5}$$

Owing to the evidence $\int q_{\phi_k}\cdot\nabla_{\phi_k}^2q_{\phi_k}\Big|_{\phi_k=\phi_k^\tau}dx_k=0$, Eq. (C3) yields

$$\mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right)=\frac{1}{2}\Delta\phi_k^{\mathrm{T}}F_{\phi_k}\Delta\phi_k+O\left(\Delta\phi_k^3\right),\ F_{\phi_k}\approx\nabla_{\phi_k}^2\mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right).\tag{C6}$$

Similar to (B3), there is

$$\mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right)=-\frac{1}{2}\left(\log|P_k^\tau|-\log|P_k|+(\hat{x}_k^\tau-\hat{x}_k)^{\mathrm{T}}(P_k^\tau)^{-1}(\hat{x}_k^\tau-\hat{x}_k)+\mathrm{Tr}\left((P_k^\tau)^{-1}P_k\right)-\mathrm{Tr}(I_{n_x\times n_x})\right).\tag{C7}$$

Compute the second-order partial derivative of (C7) with respect to $\phi_k=\{\hat{x}_k,P_k\}$, we obtain

$$\begin{cases}F_{\hat{x}_k}=\nabla_{\hat{x}_k}^2\mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right)=\dfrac{\partial^2}{2\partial{\hat{x}_k}^2}\left((\hat{x}_k^\tau-\hat{x}_k)^{\mathrm{T}}(P_k^\tau)^{-1}(\hat{x}_k^\tau-\hat{x}_k)\right)=(P_k^\tau)^{-1},\\F_{P_k}=\nabla_{P_k}^2\mathbb{D}_{KL}\left(q_{\phi_k}\middle|q_{\phi_k^\tau}\right)=\dfrac{\partial^2}{2\partial{P_k}^2}\left(\mathrm{Tr}\left((P_k^\tau)^{-1}P_k\right)-\log|P_k|\right)=-\dfrac{\partial^2\log|P_k|}{2\partial{P_k}^2}\\\qquad=-\dfrac{\partial}{2\partial P_k}\left(2P_{k|k-1}^{-1}-P_k^{-1}\circ I\right)=\dfrac{\partial P_k^{-1}}{\partial P_k}+\dfrac{\partial P_k^{-1}}{2\partial P_k}\circ I,\end{cases}\tag{C8}$$

where $\circ$ denotes Hadamard product. Computation the operator $\left(\partial P_k^{-1}/2\partial P_k\right) \circ I$ in (C8) shows that only the non-diagonal elements are left. Thus, F_{P_k} can be approximated as follows:

$$F_{P_k} \approx \frac{\partial P_k^{-1}}{2\partial P_k} = \frac{1}{2} P_k^{-1} \otimes P_k^{-1}. \tag{C9}$$

It follows that the iterative optimization yields that the final expression of (39).

Appendix D Reparameterization trick

Let us assume that the true posteriors are approximately Gaussian with an approximately diagonal covariance. In this case, we can let the variational approximate posteriors be multivariate Gaussians with a diagonal covariance structure:

$$\log q_{\phi_k}\left(x_k | y_{1:k}\right) = \log \mathcal{N}\left(x_k; \hat{x}_k, P_k\right), \tag{D1}$$

where $\hat{x}_k$ and P_k are yet unspecified functions of $y_{1:k}$. As they are Gaussian, we can reparameterize the variational approximate posteriors by

$$\tilde{x}_k = \hat{x}_k + P_k \odot \zeta, \quad \zeta \sim \mathcal{N}\left(0, I\right), \tag{D2}$$

where $\odot$ denotes an element-wise product. Then, according to the Monte-Carlo method, Eq. (45) yields

$$\mathbb{E}_{q_{\phi_k}}\left[g\left(x\right)\right] = \frac{1}{S}\sum_{s=1}^{S} g\left(\hat{x}_k + P_k \odot \zeta^{(s)}\right), \quad \zeta^{(s)} \sim \mathcal{N}\left(0, I\right). \tag{D3}$$

Through this reparameterization trick, the derivatives of the Monte-Carlo estimator (D3) with respect to the parameters $\hat{x}_k$ and P_k can be calculated. Therefore, the resulting estimator of (44) is as follows:

$$\begin{cases} \mathbb{E}_{q_{\phi_k}}\left[g \cdot \nabla_{\hat{x}_k} \log q_{\phi_k}\right] = \dfrac{1}{S}\displaystyle\sum_{s=1}^{S} g \cdot \nabla_{\hat{x}_k} \log q_{\phi_k} = -\frac{1}{S}\sum_{s=1}^{S} g \cdot \zeta^{(s)}, \\ \mathbb{E}_{q_{\phi_k}}\left[g \cdot \nabla_{P_k} \log q_{\phi_k}\right] = -\dfrac{1}{2S}\displaystyle\sum_{s=1}^{S} g \cdot P_k^{-1} \cdot \left(I - \zeta^{(s)}\left(\zeta^{(s)}\right)^{\mathrm{T}} P_k^{-1}\right). \end{cases} \tag{D4}$$