Distributionally Robust State Estimation for Linear Systems Subject to Uncertainty and Outlier

Shixiong Wang, *Student Member, IEEE*, and Zhi-Sheng Ye, *Senior Member, IEEE*

***Abstract*—Parameter uncertainties and measurement outliers unavoidably exist in a real linear system. Such uncertainties and outliers make the true joint state-measurement distributions (induced by the true system model) deviate from the nominal ones (induced by the nominal system model) so that the performance of the optimal state estimator designed for the nominal model becomes unsatisfactory or even unacceptable in practice. The challenges are to quantitatively describe the uncertainties in the model and the outliers in the measurements, and then robustify the estimator in a right way. This article studies a distributionally robust state estimation framework for linear systems subject to parameter uncertainties and measurement outliers. It utilizes a family of distributions near the nominal one to implicitly describe the uncertainties and outliers, and the robust state estimation in the worst case is made over the least-favorable distribution. The advantages of the presented framework include: 1) it only uses a few scalars to parameterize the method and does not require the structural information of uncertainties; 2) it generalizes several classical filters (e.g., the fading Kalman filter, risk-sensitive Kalman filter, relative-entropy Kalman filter, outlier-insensitive Kalman filters) into a unified framework. We show that the distributionally robust state estimation problem can be reformulated into a linear semi-definite program and in some special cases it can be analytically solved. Comprehensive comparisons with existing state estimation frameworks that are insensitive to parameter uncertainties and measurement outliers are also conducted.**

***Index Terms*—Distributionally robust, state estimation, linear system, parameter uncertainty, measurement outlier, semi-definite programming.**

I. Introduction

STATE estimation for linear systems is an important and active research topic in many fields such as target tracking, power systems, geodesy, control and automation (e.g., robotics), and astronautics (e.g., satellite attitude determination). When the noises are Gaussian, and the system matrices and the statistical properties (i.e., mean and covariance) of the noises are exactly known, the renowned Kalman filter offers the optimal estimate of the state given the measurement sequence in the sense of minimum mean square error [1] and unbiased minimum variance [2]. In most applications, however, there are uncertainties in the parameters (i.e., the system matrices and the statistical properties of the noises) of state transition and state observation models. The uncertainties might significantly deteriorate the performance of the canonical Kalman filter [3], or even cause divergence [2], [4], when the nominal values of the parameters deviate from the true ones. Even worse, the Kalman filter is sensitive to possible outliers in the measurements. There is a large body of literature on coping with uncertainties in the parameters and outliers in the measurements, leading to two streams of research.

The first stream focuses on parameter uncertainties in state estimation of linear systems. The earliest solutions include the fading (a.k.a. fading-memory) Kalman filter [4], [5], the finite horizon memory filters [6, Section V] especially the UFIR filter [7], the risk-sensitive (a.k.a. exponential-cost) Kalman filter [6, Section IV], [8], the set-valued Kalman filter [9], the $\mathcal{H}_\infty$ filter [6], [10], the adaptive Kalman filter [11]–[14], and their extensions. Comprehensive reviews and comparisons of these methods can be found in [5], [7], [12], [15], [16]. Later solutions contain the multiple-model methods which handle the case when the system modes are assumed to be multiple [17], [18], and the unknown-input filters designed for systems that have uncertain inputs [19]–[22]. Later on, robust filters that are insensitive to parameter uncertainties are introduced. They try to minimize/limit the worst-case estimation error and the uncertainties are modelled in different ways. Remarkable frameworks include the Sayed's norm-constrained filter [16], the stochastic-parameter filter [23]–[25], the relative-entropy Kalman filter [26], the τ-divergence Kalman filter [27], the Wasserstein Kalman filter [28], and the moments-based distributionally robust state estimator [29].

The second stream of research deals with outlier-insensitive state estimation. The earliest solution is the Gaussian-sum Kalman filter which approximates non-Gaussian noise distribution by a Gaussian sum [3], [30]. In order to lower the computation burden, two categories of methods are introduced afterwards. The first category uses heavy-tailed distributions for the noises which are inherently outlier-aware [31]–[33]. The second category contains the M-estimation-based Kalman filters [34]–[37]. They are designed to identify outliers and then take actions to remove/attenuate them, by leveraging various influence functions [38], [39]. A notable extension for M-estimation-based Kalman filtering is introduced in [40], which jointly estimates an unknown-input existing in both the system dynamics and the measurement dynamics.

Manuscript received May 23, 2021; revised September 13, 2021 and November 2, 2021; accepted December 16, 2021. Date of publication December 20, 2021; date of current version January 21, 2022. The associate editor coordinating the review of this manuscript and approving it for publication Dr. Monica F. Bugallo. This work was supported by the Natural Science Foundation of China under Grant 72071138, and in part by the Singapore MOE Tier 1 Grant R-266-000-145-114. *(Corresponding author: Zhi-Sheng Ye.)*

The authors are with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576, Singapore (e-mail: s.wang@u.nus.edu; yez@nus.edu.sg).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSP.2021.3136804, provided by the authors.

Digital Object Identifier 10.1109/TSP.2021.3136804

However, up to date, there does not exist a robust state estimation method that is able to address both parameter uncertainties and measurement outliers. Besides, a unified viewpoint to understand the various existing methods is lacking. Therefore, in this article,

1) We propose a distributionally robust state estimation framework for linear systems subject to parameter uncertainties and measurement outliers. It uses a family of distributions to describe the parameter uncertainties and measurement outliers, and the robust state estimation is made over the least-favorable distribution.
2) We show that the proposed framework generalizes several existing estimation methodologies, including the fading Kalman filter, the Student's t Kalman filter, the risk-sensitive Kalman filter, the M-estimation-based Kalman filters, the relative-entropy Kalman filter, the τ-divergence Kalman filter, and the Wasserstein Kalman filter.
3) We show that the proposed distributionally robust state estimation problem can be reformulated into a linear semi-definite program and in some special cases it can be analytically (i.e., efficiently) solved.
4) Comprehensive comparisons and discussions on the mentioned state-of-the-art frameworks will be made in Section VI, by comparing with the newly proposed distributionally robust estimator.

Notations. $\mathbb{R}^d$ denotes the d-dimensional Euclidean space. $\mathbb{E}_{\mathbb{P}}[\cdot]$ denotes the expectation operator of a random variable/vector/matrix with respect to the distribution $\mathbb{P}$ (the subscript $\mathbb{P}$ will be dropped when there is no ambiguity). Let $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ define a d-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathcal{D}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote any generic d-dimensional distribution. Let $\boldsymbol{Y}_k$ denote the measurement sequence up to and including time k, i.e, $\boldsymbol{Y}_k := \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$. Let $\boldsymbol{I}$ and $\boldsymbol{0}$ denote an identity and a null matrix with appropriate dimensions, respectively. We use $\boldsymbol{M}^T$ to denote the transpose of the matrix $\boldsymbol{M}$, and $\mathrm{Tr}[\boldsymbol{M}]$ its trace when $\boldsymbol{M}$ is square. We use $\mathbb{P}_{\boldsymbol{x},\cdots}(\boldsymbol{x}, \cdots)$ to denote the joint/conditional/marginal distribution of $\{\boldsymbol{x}, \cdots\}$. Whenever no confusion is caused, we drop the subscript of $\mathbb{P}$ and write it as $\mathbb{P}(\boldsymbol{x}, \cdots)$. Let $\mathbb{S}^d$ denote the set of all d-dimensional symmetric matrices in $\mathbb{R}^{d\times d}$, and $\mathbb{S}^d_+$ (resp. $\mathbb{S}^d_{++}$) of all d-dimensional symmetric positive semi-definite (resp. positive definite) matrices in $\mathbb{S}^d$. If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^d$, $\boldsymbol{A} \succeq \boldsymbol{B}$ (resp. $\boldsymbol{A} \succ \boldsymbol{B}$) indicates that $\boldsymbol{A} - \boldsymbol{B} \in \mathbb{S}^d_+$ (resp. $\boldsymbol{A} - \boldsymbol{B} \in \mathbb{S}^d_{++}$). If $\boldsymbol{S} \in \mathbb{S}^d_+$, let $\boldsymbol{S}^{1/2}$ be the square root of $\boldsymbol{S}$ (i.e., $\boldsymbol{S}^{1/2}\boldsymbol{S}^{1/2} = \boldsymbol{S}$). To avoid notation clutter, an ellipsis in a bracket means a copy of the content in the immediately previous bracket (e.g., $[\mathcal{E}][\cdots] := [\mathcal{E}][\mathcal{E}]$ when an expression $\mathcal{E}$ is long).

II. Preliminaries on Distributional Robustness

The original concept of distributional robustness stemmed from the statistical game theory (recall the mixed strategy) [41], the inventory problem [42], and the Huber's outlier-insensitive robust statistics [43]. It is now known for distributionally robust optimization and popular in operations research [44], machine learning [45], [46], and systems control [47]. Let $\boldsymbol{x} \in \mathcal{X}$ denote the decision vector and $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ the random parameter vector associated with the optimization problem $\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ where $f(\cdot, \cdot)$ is the objective function. In practice, however, we do not exactly know the true distribution $\mathbb{P}_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$. This motivates us to assume that the true distribution $\mathbb{P}_{\boldsymbol{\xi}}$ lies in a family of distributions $\mathcal{F}$ and find the worst-case robust optimality, i.e.,

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}_{\boldsymbol{\xi}}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}[f(\boldsymbol{x}, \boldsymbol{\xi})], \tag{1}$$

where $\mathcal{F}$ is called the **ambiguity set**, given as

$$\mathcal{F} = \left\{ \mathbb{P}_{\boldsymbol{\xi}} \,\middle|\, \begin{array}{l} \boldsymbol{\xi} \sim \mathbb{P}_{\boldsymbol{\xi}} \\ \mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\xi} \in \boldsymbol{\Xi}) = 1 \\ \text{other requirements} \end{array} \right\}.$$

The said "other requirements" describes the relation between the ambiguity set and the nominal distribution $\bar{\mathbb{P}}_{\boldsymbol{\xi}}$. Typically, the ambiguity set is constructed as a ball centered at the nominal distribution, e.g., under the Wasserstein distance [48]

$$\mathcal{F}_W = \left\{ \mathbb{P}_{\boldsymbol{\xi}} \,\middle|\, \begin{array}{l} \boldsymbol{\xi} \sim \mathbb{P}_{\boldsymbol{\xi}} \\ \mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\xi} \in \boldsymbol{\Xi}) = 1 \\ \mathrm{W}(\mathbb{P}_{\boldsymbol{\xi}}, \bar{\mathbb{P}}_{\boldsymbol{\xi}}) \leq \theta \end{array} \right\}, \tag{2}$$

where $\mathrm{W}(\cdot, \cdot)$ defines the Wasserstein distance. Intuitively, although we do not know the true distribution, we assume that the true governing distribution is not far away from the nominal one. The radius θ of the ball adjusts our trust level towards the nominal distribution. Other possible construction methods include the Kullback-Leibler divergence [49], the τ-divergence [27], the ϕ-divergence [49] (a.k.a. f-divergence), the moment-based ambiguity set [50], etc.

If $\boldsymbol{x}^*$ and $\mathbb{P}^*_{\boldsymbol{\xi}}$ solve the distributionally robust optimization problem (1), we term $\boldsymbol{x}^*$ as the worst-case robust solution and $\mathbb{P}^*_{\boldsymbol{\xi}}$ as the least-favorable (a.k.a. worst-case) distribution.

III. Problem Formulation

We are concerned with estimating the hidden state vector $\boldsymbol{x}_k$ of a linear Markov system [2], [51], [52]

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1}, \\ \boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k, \end{cases} \tag{3}$$

where k is the discrete time index; $\boldsymbol{x}_k \in \mathbb{R}^n$ is the state vector; $\boldsymbol{y}_k \in \mathbb{R}^m$ is the measurement vector; $\boldsymbol{w}_{k-1} \in \mathbb{R}^p$, $\boldsymbol{v}_k \in \mathbb{R}^m$ are the process noise and measurement noise, respectively. Typically, the linear system (3) is assumed to have the following properties [2], [51], [52]: 1) For every k, $\boldsymbol{x}_k$, $\boldsymbol{y}_k$, $\boldsymbol{w}_k$, and $\boldsymbol{v}_k$ have finite second moments; 2) $\boldsymbol{x}_0 \sim \mathcal{N}_n(\bar{\boldsymbol{x}}_0, \boldsymbol{M}_0)$, $\boldsymbol{w}_k \sim \mathcal{N}_p(\boldsymbol{\mu}^{\boldsymbol{w}}_k, \boldsymbol{Q}_k)$, and $\boldsymbol{v}_k \sim \mathcal{N}_m(\boldsymbol{\mu}^{\boldsymbol{v}}_k, \boldsymbol{R}_k)$. Besides, $\boldsymbol{\mu}^{\boldsymbol{w}}_k$, and $\boldsymbol{\mu}^{\boldsymbol{v}}_k$ are exactly known and typically zero-valued; 3) For every $j \neq k$, $\boldsymbol{w}_k$ and $\boldsymbol{x}_0$ are uncorrelated, so are $\boldsymbol{v}_k$ and $\boldsymbol{x}_0$, $\boldsymbol{w}_k$ and $\boldsymbol{w}_j$, and $\boldsymbol{v}_k$ and $\boldsymbol{v}_j$. For every k, j, $\boldsymbol{v}_k$ and $\boldsymbol{w}_j$ are uncorrelated; 4) $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, $\boldsymbol{F}_{k-1}$, $\boldsymbol{G}_{k-1}$, and $\boldsymbol{H}_k$ are exactly known and have finite matrix norms.

The nominal system (3) defines two discrete time stochastic processes $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$, $k = 1, 2, \cdots$. Let $\mathcal{H}_{\boldsymbol{Y}_k}$ denote the collection of all possible linear combinations of $\{\boldsymbol{1}, \boldsymbol{Y}_k\}$ and $\mathcal{H}'_{\boldsymbol{Y}_k}$ denote the collection of all second-moment-finite Borel measurable functions of $\{\boldsymbol{1}, \boldsymbol{Y}_k\}$. Suppose the nominal joint state-measurement distribution defined by the nominal system model (3) is $\bar{\mathbb{P}}(\boldsymbol{x}_k, \boldsymbol{Y}_k)$. We would like to solve the following

optimization problem

$$\min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{Y}_k}} \mathbb{E}_{\bar{\mathbb{P}}(\boldsymbol{x}_k,\boldsymbol{Y}_k)}[\boldsymbol{x}_k-\boldsymbol{\phi}(\boldsymbol{Y}_k)][\boldsymbol{x}_k-\boldsymbol{\phi}(\boldsymbol{Y}_k)]^T, \tag{4}$$

where $\boldsymbol{\phi}(\cdot)$ is called an optimal estimator. The optimal estimate of $\boldsymbol{x}_k$ in this minimum mean square error sense is $\mathbb{E}(\boldsymbol{x}_k|\boldsymbol{Y}_k)\in\mathcal{H}'_{\boldsymbol{Y}_k}$. In particular, if $\bar{\mathbb{P}}(\boldsymbol{x}_k,\boldsymbol{Y}_k)$ is jointly Gaussian, $\mathbb{E}(\boldsymbol{x}_k|\boldsymbol{Y}_k)$ has a linear form, i.e., $\mathbb{E}(\boldsymbol{x}_k|\boldsymbol{Y}_k)\in\mathcal{H}_{\boldsymbol{Y}_k}$. Nice properties (e.g., linearity, Gaussianity) of the nominal system (3) produce a beautiful solution to (4), i.e., the Kalman filter. However, in general, problem (4) is not always easy to solve if the involved distribution $\bar{\mathbb{P}}(\boldsymbol{x}_k,\boldsymbol{Y}_k)$ is not Gaussian [53].

Remark 1: The objective in (4) is a positive semi-definite matrix. In fact, minimizing a matrix objective is equivalent to minimizing its trace [54], [55]. Note that $\min_{\boldsymbol{X}\in\mathcal{X}}\boldsymbol{X}$ and $\min_{\boldsymbol{X}\in\mathcal{X}}\mathrm{Tr}[\boldsymbol{X}]$ over a convex and compact matrix set $\mathcal{X}$ have the same matrix-valued solution $\boldsymbol{X}^*$ because the trace operator is monotonically increasing. □

If the underlying system dynamics deviates from the nominal model (3), the true joint state-measurement distribution $\mathbb{P}(\boldsymbol{x}_k,\boldsymbol{Y}_k)$ will more or less diverge from the nominal $\bar{\mathbb{P}}(\boldsymbol{x}_k,\boldsymbol{Y}_k)$. In this scenario, we aim to find a robust state estimation solution that is insensitive to the deviation. Inspired by the distributionally robust optimization theory, we can write the distributionally robust counterpart of (4) as

$$\min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{Y}_k}} \max_{\mathbb{P}(\boldsymbol{x}_k,\boldsymbol{Y}_k)\in\mathcal{F}} \mathbb{E}_{\mathbb{P}(\boldsymbol{x}_k,\boldsymbol{Y}_k)}[\boldsymbol{x}_k-\boldsymbol{\phi}(\boldsymbol{Y}_k)][\boldsymbol{x}_k-\boldsymbol{\phi}(\boldsymbol{Y}_k)]^T, \tag{5}$$

where $\mathcal{F}$ is the associated ambiguity set constructed around the nominal distribution $\bar{\mathbb{P}}(\boldsymbol{x}_k,\boldsymbol{Y}_k)$. This worst-case optimization problem can be treated as a zero-sum statistical game [41] where the two adversarial players are the statistician who chooses the optimal estimator and the nature that chooses the uncertain, hostile distribution (i.e., one tries to lower the cost but the other to improve).

Nevertheless, directly solving the min-max problem (5) is doubtful because we prefer a recursive-type solution. Note that the optimal estimator operates along the discrete time in a recursive way [54] because state estimation is an online (i.e., time-series) problem. This also helps to reduce the calculation complexity at each time step. Thus, we instead try to solve a time-incremental [26] (i.e., one-time-step) alternative problem

$$\min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{y}_k}} \max_{\mathbb{P}(\boldsymbol{x}_k,\boldsymbol{y}_k|\boldsymbol{Y}_{k-1})\in\mathcal{F}'} \mathbb{E}_{\mathbb{P}(\cdot,\cdot|\cdot)}[\boldsymbol{x}_k-\boldsymbol{\phi}(\boldsymbol{y}_k)][\boldsymbol{x}_k-\boldsymbol{\phi}(\boldsymbol{y}_k)]^T, \tag{6}$$

where the new ambiguity set $\mathcal{F}'$ is constructed around the nominal conditional joint state-measurement distribution given the previous measurement sequence $\bar{\mathbb{P}}(\boldsymbol{x}_k,\boldsymbol{y}_k|\boldsymbol{Y}_{k-1})$. Note that in (6), the space of $\boldsymbol{\phi}(\cdot)$ is only defined by $\boldsymbol{y}_k$ instead of $\boldsymbol{Y}_k$. In order to solve (6), we need to: 1) design proper forms of the associated ambiguity set $\mathcal{F}'$ so that both the parameter uncertainties and measurement outliers can be taken into consideration, and 2) find the explicit optimization equivalent(s) of (6) so that it can be efficiently solved.

Therefore, we are inspired to **first** study a distributionally robust Bayesian estimation problem

$$\min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{y}}} \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''} \mathbb{E}_{\mathbb{P}(\cdot,\cdot)}[\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})][\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})]^T \tag{7}$$

subject to the nominal prior state distribution $\bar{\mathbb{P}}(\boldsymbol{x})$, the nominal conditional measurement distribution $\bar{\mathbb{P}}(\boldsymbol{y}|\boldsymbol{x})$, a properly constructed ambiguity set $\mathcal{F}''$, and the linear measurement equation

$$\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}+\boldsymbol{v}, \tag{8}$$

where $\boldsymbol{x},\boldsymbol{y},\boldsymbol{v}$ are second-moment-finite with appropriate dimensions and distributions. The subscript k (i.e., discrete time index) is dropped to avoid notation clutter. **Then**, by identifying the joint distribution of $(\boldsymbol{x}_k,\boldsymbol{y}_k)$ conditioned on $\boldsymbol{Y}_{k-1}$, we can solve (6).

In order to make the problem (7) tractable, we assume that the nominal $\bar{\mathbb{P}}_{\boldsymbol{x}}$ and $\bar{\mathbb{P}}_{\boldsymbol{v}}$ are Gaussian. In other words, no matter what the true distributions $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{P}_{\boldsymbol{v}}$ are, we use Gaussian distributions to approximate them. The Gaussian approximation is popular in state estimation community, especially for nonlinear systems. For instance, recall the cubature Kalman filter [56], the unscented Kalman filter [57], etc. Besides, the Gaussian distribution has the following properties, which adapt into our worst-case robust perspective.

1) The Gaussian distribution admits maximum entropy (i.e., maximum degree of indeterminacy) among all distributions with given/fixed first- and second-order moments [58].
2) Concerning a linear measurement system $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}+\boldsymbol{v}$, if the state $\boldsymbol{x}$ is Gaussian, then among all noise distributions with bounded variance for $\boldsymbol{v}$, the Gaussian minimizes the mutual information between the state $\boldsymbol{x}$ and the measurement $\boldsymbol{y}$. Namely, the Gaussian noise makes the measurement least informative to estimate the state [53], [59].
3) Concerning the linear measurement system above, if the noise $\boldsymbol{v}$ is Gaussian, then among all state distributions with bounded variance for $\boldsymbol{x}$, the Gaussian maximizes the minimum mean square error. Namely, the Gaussian state is most difficult to estimate [53], [60].

The third reason to make the Gaussianity assumption is that the Wasserstein metric and the Kullback–Leibler divergence for Gaussian distributions admit closed-form expressions.

IV. Distributionally Robust Bayesian Estimation

With linear measurement relation (8), the joint state-measurement distribution $\mathbb{P}(\boldsymbol{x},\boldsymbol{y})$ can be determined by (specifically, linearly shifted from) $\mathbb{P}(\boldsymbol{x},\boldsymbol{v})$ which has marginals $\mathbb{P}(\boldsymbol{x})$ and $\mathbb{P}(\boldsymbol{v})$. In such a situation, it is reasonable and common to assume that the state $\boldsymbol{x}$ is independent of the measurement noise $\boldsymbol{v}$. As a result, we have $\mathbb{P}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})=\mathbb{P}_{\boldsymbol{x},\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})=\mathbb{P}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x})=\mathbb{P}_{\boldsymbol{v}|\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}|\boldsymbol{x})\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x})=\mathbb{P}_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x})$. Therefore, $\mathbb{P}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})=\mathbb{P}_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})$.

To solve the primal problem (7), we are required to identify the least-favorable distribution from the ambiguity set $\mathcal{F}''$. However, it depends on the specific choice of the estimator $\boldsymbol{\phi}(\cdot)$. Therefore, we can alternatively try to solve the dual problem of (7) first, i.e.,

$$\max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''} \min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{y}}} \mathbb{E}_{\mathbb{P}(\cdot,\cdot)}[\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})][\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})]^T. \tag{9}$$

However, the dual problem (9) and the primal problem (7) are not guaranteed to be equivalent. In general, by the weak duality,

we have $\max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''}\min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{y}}}\mathbb{E}_{\mathbb{P}(\cdot,\cdot)}[\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})][\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})]^T \preceq \min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{y}}}\max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''}\mathbb{E}_{\mathbb{P}(\cdot,\cdot)}[\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})][\boldsymbol{x}-\boldsymbol{\phi}(\boldsymbol{y})]^T$. The equality stands only when the strong duality holds which is not always the case. The dual problem is easier to solve because for every $\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''$, we can find the associated optimal estimator. We first study the optimal estimator for the nominal case.

Theorem 1: Suppose $\boldsymbol{x}\sim\mathcal{N}_n(\bar{\boldsymbol{x}},\boldsymbol{M})$ nominally, $\boldsymbol{x}$ is independent of $\boldsymbol{v}$, all involved densities exist, and all involved integration and differentiation are interchangeable (i.e., densities are twice continuously differentiable). Let $\boldsymbol{s}:=\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}}$ denote the innovation vector, $\boldsymbol{S}$ the associated covariance, and $\boldsymbol{\mu}:=\boldsymbol{S}^{-1/2}\boldsymbol{s}$ the diagonalized and normalized innovation. Then for any possible nominal joint state-measurement distribution $\bar{\mathbb{P}}(\boldsymbol{x},\boldsymbol{y})$, the optimal estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$, i.e., $\mathbb{E}(\boldsymbol{x}|\boldsymbol{y})$, is

$$\hat{\boldsymbol{x}}=\bar{\boldsymbol{x}}+\boldsymbol{M}\boldsymbol{H}^T\boldsymbol{S}^{-1/2}\left[-\frac{d}{d\boldsymbol{\mu}}\ln p(\boldsymbol{\mu})\right], \tag{10}$$

and the conditional covariance of the estimation error given $\boldsymbol{y}$, i.e., $\boldsymbol{P}_{\boldsymbol{x}|\boldsymbol{y}}:=\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(\hat{\boldsymbol{x}}-\boldsymbol{x})(\hat{\boldsymbol{x}}-\boldsymbol{x})^T$, is

$$\boldsymbol{P}_{\boldsymbol{x}|\boldsymbol{y}}=\boldsymbol{M}-\boldsymbol{M}\boldsymbol{H}^T\boldsymbol{S}^{-1/2}\left[-\frac{d}{d\boldsymbol{\mu}}\ln p(\boldsymbol{\mu})\right][\cdots]^T\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{M}, \tag{11}$$

where

$$p(\boldsymbol{\mu})=p_{\boldsymbol{\mu}}(\boldsymbol{\mu})=p_{\boldsymbol{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu}+\boldsymbol{H}\bar{\boldsymbol{x}})\cdot\det(\boldsymbol{S}^{1/2}) \tag{12}$$

is the density of $\boldsymbol{\mu}$, $p_{\boldsymbol{y}}(\cdot)$ is the density of $\boldsymbol{y}$, and $\det(\cdot)$ denotes the determinant of a matrix.

Proof: See Appendix A. □

Theorem 1 reveals the benefit of the Gaussianity assumption of $\bar{\mathbb{P}}_{\boldsymbol{x}}$. Specifically, without the Gaussianity assumption, we cannot have the closed form of $\hat{\boldsymbol{x}}$ as in (10).

Corollary 1: With the posterior estimate $\hat{\boldsymbol{x}}$, the posterior estimation error covariance is given by $\boldsymbol{P}:=\mathbb{E}_{\boldsymbol{y}}\boldsymbol{P}_{\boldsymbol{x}|\boldsymbol{y}}$ and

$$\boldsymbol{P}=\boldsymbol{M}-\boldsymbol{M}\boldsymbol{H}^T\boldsymbol{S}^{-1/2}\mathbb{E}\left[-\frac{d^2}{d\boldsymbol{\mu}\boldsymbol{\mu}^T}\ln p(\boldsymbol{\mu})\right]\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{M}. \tag{13}$$

Proof: See Appendix B. □

From $\boldsymbol{\mu}=\boldsymbol{S}^{-1/2}(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})=\boldsymbol{S}^{-1/2}[\boldsymbol{H}(\boldsymbol{x}-\bar{\boldsymbol{x}})+\boldsymbol{v}]$, we know that $\mathbb{E}\boldsymbol{\mu}=\boldsymbol{0}$ and $\mathbb{E}\boldsymbol{\mu}\boldsymbol{\mu}^T=\boldsymbol{I}$. Thus, if $\boldsymbol{x}$ and $\boldsymbol{v}$ were all normally distributed, $\boldsymbol{p}_{\boldsymbol{\mu}}(\cdot)$ would be standard Gaussian because the independence between $\boldsymbol{x}$ and $\boldsymbol{v}$ has already been assumed. Specifically, for every $i\neq j$, $\mathbb{E}\mu_i=\mathbb{E}\mu_j=0$, $\mathbb{E}\mu_i^2=\mathbb{E}\mu_j^2=1$, and $\mathbb{E}\mu_i\mu_j=0$. Therefore, we have

$$\mathbb{E}\left[-\frac{d^2}{d\boldsymbol{\mu}\boldsymbol{\mu}^T}\ln p(\boldsymbol{\mu})\right]=\boldsymbol{I}\cdot\mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right]. \tag{14}$$

Note that the entry-wise $p_\mu(\cdot)$ is different from the joint $p_{\boldsymbol{\mu}}(\cdot)$ and we have $p(\boldsymbol{\mu})=\prod_i p(\mu_i)$. For a nominal Gaussian $p_\mu(\cdot)$, we identify that $-\frac{d}{d\mu}\ln p(\mu)$ is the score function (i.e., maximum likelihood estimator of the mean; see also Appendix C) of the distribution $p_\mu(\cdot)$ and $\mathbb{E}[-\frac{d^2}{d\mu^2}\ln p(\mu)]$ the associated Fisher information (whose reciprocal gives the lower bound of the asymptotic variance obtainable by the mean maximum likelihood estimator). Equation (14) is attractive since it allows us to only study a univariate problem rather than a multivariate one. This motivated us to study the normalized and diagnalized innovation $\boldsymbol{\mu}$ instead of $\boldsymbol{s}$. Hence, (13) can be simplified to

$$\boldsymbol{P}=\boldsymbol{M}-\boldsymbol{M}\boldsymbol{H}^T\boldsymbol{S}^{-1}\boldsymbol{H}\boldsymbol{M}\cdot\mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right]. \tag{15}$$

By the results in Corollary 1, we can find the explicit and tractable reformulation of the dual problem (9).

Corollary 2: Suppose the true distribution of $\boldsymbol{x}$ is $\mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}},\boldsymbol{\Sigma}_{\boldsymbol{x}})$ [cf. the nominal $\mathcal{N}_n(\bar{\boldsymbol{x}},\boldsymbol{M})$ in Theorem 1]. The dual problem (9) can be reformulated as

$$\max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''}\boldsymbol{P}, \tag{16}$$

where $\mathbb{P}(\boldsymbol{x},\boldsymbol{y})$ is a possible joint distribution of $(\boldsymbol{x},\boldsymbol{y})$,

$$\boldsymbol{P}=\boldsymbol{\Sigma}_{\boldsymbol{x}}-\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^T\boldsymbol{S}^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\cdot\mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right], \tag{17}$$

and $\boldsymbol{S}$ is the covariance matrix of $\boldsymbol{s}:=\boldsymbol{y}-\boldsymbol{H}\boldsymbol{c}_x$.

Proof: This is immediate from Theorem 1, Corollary 1, and (15). In this case, $\hat{\boldsymbol{x}}=\boldsymbol{c}_x+\boldsymbol{\Sigma}_x\boldsymbol{H}^T\boldsymbol{S}^{-1/2}[-\frac{d}{d\boldsymbol{\mu}}\ln p(\boldsymbol{\mu})]$ and $\boldsymbol{\mu}:=\boldsymbol{S}^{-1/2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}})$. Since $\mathbb{E}[-\frac{d}{d\boldsymbol{\mu}}\ln p(\boldsymbol{\mu})]=-\int[p(\boldsymbol{\mu})]^{-1}\frac{dp(\boldsymbol{\mu})}{d\boldsymbol{\mu}}p(\boldsymbol{\mu})d\boldsymbol{\mu}=-\int\frac{dp(\boldsymbol{\mu})}{d\boldsymbol{\mu}}d\boldsymbol{\mu}=\boldsymbol{0}$, we have $\mathbb{E}\hat{\boldsymbol{x}}=\boldsymbol{c}_x=\mathbb{E}\boldsymbol{x}$, implying that $\hat{\boldsymbol{x}}$ is an unbiased estimate so that the minimum mean square error matrix coincides with the minimum error covariance matrix. □

To explicitly solve (16), we need to define the ambiguity set $\mathcal{F}''$. The prior state distribution is Gaussian as argued. We can construct the ambiguity set for $\mathbb{P}(\boldsymbol{x})$ as

$$\mathcal{F}_{\boldsymbol{x}}=\left\{\mathbb{P}_{\boldsymbol{x}}\;\middle|\;\begin{array}{l}\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}\\ \mathbb{P}_{\boldsymbol{x}}=\mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}},\boldsymbol{\Sigma}_{\boldsymbol{x}})\\ \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x}\in\mathbb{R}^n)=1\\ \Delta(\mathbb{P}_{\boldsymbol{x}},\bar{\mathbb{P}}_{\boldsymbol{x}})\leq\theta\end{array}\right\}.$$

where $\bar{\mathbb{P}}_{\boldsymbol{x}}$ is the nominal Gaussian distribution of $\boldsymbol{x}$ [i.e., $\mathcal{N}_n(\bar{\boldsymbol{x}},\boldsymbol{M})$ in Theorem 1], $\Delta(\cdot,\cdot)$ is a statistical metric (e.g., Wasserstein metric) or divergence (e.g., Kullback–Leibler divergence), and $\theta\in\mathbb{R}_+$ is the radius to control the scale and conservativeness of the set. The larger the θ, the more conservative the robust estimation is. Specially, the ambiguity set for $\mathbb{P}(\boldsymbol{x})$ could be one of the follows.

1) **Kullback–Leibler divergence** (KL divergence).

$$\mathcal{F}_{\boldsymbol{x}}=\left\{\mathbb{P}_{\boldsymbol{x}}\;\middle|\;\begin{array}{l}\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}\\ \mathbb{P}_{\boldsymbol{x}}=\mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}},\boldsymbol{\Sigma}_{\boldsymbol{x}})\\ \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x}\in\mathbb{R}^n)=1\\ \mathrm{KL}(\mathbb{P}_{\boldsymbol{x}}\|\bar{\mathbb{P}}_{\boldsymbol{x}})\leq\theta_{\boldsymbol{x}}\end{array}\right\}, \tag{18}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the KL divergence and under Gaussianity assumption, $\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}}\|\bar{\mathbb{P}}_{\boldsymbol{x}})=\frac{1}{2}[\|\boldsymbol{c}_{\boldsymbol{x}}-\bar{\boldsymbol{x}}\|^2_{\boldsymbol{M}^{-1}}+\mathrm{Tr}[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}-\boldsymbol{I}]-\ln\det(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}})]$ [26]. Note that the explicit expression for any two multivariate distributions does not always exist. Only for Gaussians, the above equality holds. Extensions and generalizations for the KL divergence include the τ-divergence [27], the ϕ-divergence (a.k.a. f-divergence) [49], etc. They all contain the KL divergence as a special case.

2) **Wasserstein distance**.

$$\mathcal{F}_{\boldsymbol{x}} = \left\{ \mathbb{P}_{\boldsymbol{x}} \left| \begin{array}{l} \boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}} \\ \mathbb{P}_{\boldsymbol{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \\ \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x} \in \mathbb{R}^n) = 1 \\ \mathrm{W}(\mathbb{P}_{\boldsymbol{x}}, \bar{\mathbb{P}}_{\boldsymbol{x}}) \leq \theta_{\boldsymbol{x}} \end{array} \right. \right\}, \tag{19}$$

where $\mathrm{W}(\cdot,\cdot)$ denotes the Wasserstein metric and under Gaussianity assumption, the type-2 Wasserstein distance is given as $\mathrm{W}(\mathbb{P}_{\boldsymbol{x}}, \bar{\mathbb{P}}_{\boldsymbol{x}}) = \sqrt{\|\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2 + \mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}]}$ [28], [53]. Note also that the explicit expression for any two multivariate distributions does not always exist. Only for Gaussians, the above equality holds.

3) **Moment-based set** [50].

$$\mathcal{F}_{\boldsymbol{x}} = \left\{ \mathbb{P}_{\boldsymbol{x}} \left| \begin{array}{l} \boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}} \\ \mathbb{P}_{\boldsymbol{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \\ \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x} \in \mathbb{R}^n) = 1 \\ [\mathbb{E}(\boldsymbol{x}) - \bar{\boldsymbol{x}}]^T \boldsymbol{M}^{-1} [\mathbb{E}(\boldsymbol{x}) - \bar{\boldsymbol{x}}] \leq \theta_{3,\boldsymbol{x}} \\ \mathbb{E}(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^T \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \mathbb{E}(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^T \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \end{array} \right. \right\}$$

$$= \left\{ \mathbb{P}_{\boldsymbol{x}} \left| \begin{array}{l} \boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}} \\ \mathbb{P}_{\boldsymbol{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \\ \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{x} \in \mathbb{R}^n) = 1 \\ [\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}]^T \boldsymbol{M}^{-1} [\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}] \leq \theta_{3,\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} + (\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})^T \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} + (\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})^T \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \end{array} \right. \right\}. \tag{20}$$

As we can see, in general, we need to use three parameters to define a moment-based ambiguity set, $\theta_3 \geq 0$ and $\theta_2 \geq 1 \geq \theta_1 \geq 0$.

Suppose the nominal distribution of the measurement noise $\boldsymbol{v}$, by the Gaussianity assumption, is $\bar{\mathbb{P}}(\boldsymbol{v}) := \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{R})$. The ambiguity set for $\mathbb{P}(\boldsymbol{v})$ can be one of the follows.

1) **Kullback–Leibler divergence** (KL divergence).

$$\mathcal{F}_{\boldsymbol{v}} = \left\{ \mathbb{P}_{\boldsymbol{v}} \left| \begin{array}{l} \boldsymbol{v} \sim \mathbb{P}_{\boldsymbol{v}} \\ \mathbb{P}_{\boldsymbol{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \\ \mathbb{P}_{\boldsymbol{v}}(\boldsymbol{v} \in \mathbb{R}^m) = 1 \\ \mathrm{KL}(\mathbb{P}_{\boldsymbol{v}} \| \bar{\mathbb{P}}_{\boldsymbol{v}}) \leq \theta_{\boldsymbol{v}} \end{array} \right. \right\}. \tag{21}$$

2) **Wasserstein distance**.

$$\mathcal{F}_{\boldsymbol{v}} = \left\{ \mathbb{P}_{\boldsymbol{v}} \left| \begin{array}{l} \boldsymbol{v} \sim \mathbb{P}_{\boldsymbol{v}} \\ \mathbb{P}_{\boldsymbol{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \\ \mathbb{P}_{\boldsymbol{v}}(\boldsymbol{v} \in \mathbb{R}^m) = 1 \\ \mathrm{W}(\mathbb{P}_{\boldsymbol{v}}, \bar{\mathbb{P}}_{\boldsymbol{v}}) \leq \theta_{\boldsymbol{v}} \end{array} \right. \right\}. \tag{22}$$

3) **Moment-based set**.

$$\mathcal{F}_{\boldsymbol{v}} = \left\{ \mathbb{P}_{\boldsymbol{v}} \left| \begin{array}{l} \boldsymbol{v} \sim \mathbb{P}_{\boldsymbol{v}} \\ \mathbb{P}_{\boldsymbol{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \\ \mathbb{P}_{\boldsymbol{v}}(\boldsymbol{v} \in \mathbb{R}^m) = 1 \\ [\boldsymbol{c}_{\boldsymbol{v}} - \boldsymbol{0}]^T \boldsymbol{R}^{-1} [\boldsymbol{c}_{\boldsymbol{v}} - \boldsymbol{0}] \leq \theta_{3,\boldsymbol{v}} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} + (\boldsymbol{c}_{\boldsymbol{v}} - \boldsymbol{0})(\boldsymbol{c}_{\boldsymbol{v}} - \boldsymbol{0})^T \preceq \theta_{2,\boldsymbol{v}} \boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} + (\boldsymbol{c}_{\boldsymbol{v}} - \boldsymbol{0})(\boldsymbol{c}_{\boldsymbol{v}} - \boldsymbol{0})^T \succeq \theta_{1,\boldsymbol{v}} \boldsymbol{R} \end{array} \right. \right\}. \tag{23}$$

The explicit expressions for $\mathrm{KL}(\cdot\|\cdot)$ and $\mathrm{W}(\cdot,\cdot)$ are similar to those for $\mathbb{P}(\boldsymbol{x})$ in (18) and (19), respectively.

Given the nominal Gaussian distributions of the state $\boldsymbol{x}$ and the measurement noise $\boldsymbol{v}$, the marginal distribution of the measurement $\boldsymbol{y}$ (or equivalently, the innovation $\boldsymbol{s}$ and $\boldsymbol{\mu}$) is also Gaussian, so is the joint state-measurement distribution. However, when outliers appear in the measurement $\boldsymbol{y}$, they appear in the normalized innovation $\boldsymbol{\mu}$ (and μ) as well. That means the true distribution of $p_\mu(\cdot)$ is likely to deviate from the nominal Gaussian and, simultaneously, has a heavy tail. Let $\Phi(\cdot)$ denote the nominal standard Gaussian distribution of μ. Motivated by the M-estimation theory for outlier attenuation/rejection [43], we can construct the ambiguity set for $p(\mu)$ as one of the follows.

1) ϵ**-contamination set**.

$$\mathcal{F}_\mu = \left\{ \mathbb{P}_\mu \left| \begin{array}{l} \mu \sim \mathbb{P}_\mu \\ \mathbb{P}_\mu(\mu \in \mathbb{R}) = 1 \\ \sup_\mu \|\mathbb{P}_\mu - \Phi(\cdot)\| \leq \epsilon \\ \mathbb{P}_\mu = (1-\epsilon)\Phi + \epsilon\mathbb{H} \\ \mathbb{H}(\mu) = 1 - \mathbb{H}(-\mu) \end{array} \right. \right\}. \tag{24}$$

Note that $\sup_\mu \|\mathbb{P}_\mu - \Phi(\cdot)\| = \sup_\mu \|(1-\epsilon)\Phi + \epsilon\mathbb{H} - \Phi(\cdot)\| = \epsilon \cdot \sup_\mu \|\mathbb{H} - \Phi(\cdot)\| \leq \epsilon$ (i.e., in this case the statistical metric is instantiated as the infinity norm). The argument also holds for the total variation metric. Suppose z is an indicator and uniformly distributed in the interval [0, 1]. $\mathbb{P}_\mu = \int_z \mathbb{P}(\mu, z)dz = \Phi(\mu)\mathbb{I}_{(z\geq\epsilon)} + \mathbb{H}(\mu)\mathbb{I}_{(z\leq\epsilon)} = (1-\epsilon)\Phi + \epsilon\mathbb{H}$ where $\mathbb{I}_{(\cdot)}$ is the indicator function. Therefore, in (24), $\mathbb{P}_\mu = (1-\epsilon)\Phi + \epsilon\mathbb{H}$ means that with probability $1-\epsilon$ the measurement innovation μ (equivalently, a measurement $\boldsymbol{y}$) is from a nominal Gaussian, and with probability ϵ it is from a contamination heavy-tailed distribution $\mathbb{H}(\cdot)$ (i.e., outlier). $\mathbb{H}(\mu) = 1 - \mathbb{H}(-\mu)$ means that $\mathbb{H}(\cdot)$ is symmetric about $\mu = 0$.

2) ϵ**-normal set**.

$$\mathcal{F}_\mu = \left\{ \mathbb{P}_\mu \left| \begin{array}{l} \mu \sim \mathbb{P}_\mu \\ \mathbb{P}_\mu(\mu \in \mathbb{R}) = 1 \\ \sup_\mu \|\mathbb{P}_\mu - \Phi(\cdot)\| \leq \epsilon \\ \mathbb{P}_\mu(\mu) = 1 - \mathbb{P}_\mu(-\mu) \end{array} \right. \right\}. \tag{25}$$

Clearly, the ϵ-normal set is larger and more general than the ϵ-contamination set for the same radius ϵ. However, we usually prefer the ϵ-contamination set because: 1) it has clearer physical meaning than that of the ϵ-normal set; 2) in view of properties of real measurement data, the least-favorable distribution in (24) is more reasonable than that in the ϵ-normal set; and 3) the distributionally robust state estimator over the ϵ-contamination set is much easier to design. Other possible choice for the structure of $\mathcal{F}_\mu$ includes the p-value set [34] which is also a subset of (25), etc.

Note that the distribution of the innovation $\boldsymbol{\mu}$ is uniquely determined given the distributions of the state $\boldsymbol{x}$ and measurement noise $\boldsymbol{v}$, because we have $\boldsymbol{\mu} = \boldsymbol{S}^{-1/2}[\boldsymbol{H}(\boldsymbol{x} - \boldsymbol{c}_x) + \boldsymbol{v}]$. Thus, when we admit the ϵ-contamination/normal deviation from the nominal Gaussian distribution of $\boldsymbol{\mu}$, we implicitly admit that from the distribution(s) of $\boldsymbol{x}$ or $\boldsymbol{v}$ or both. Since the ϵ-contamination/normal deviation studied here accounts for measurement outliers, we argue that it is related to $\boldsymbol{v}$ and regardless of $\boldsymbol{x}$. However, for simplicity in problem solving, we work on $\boldsymbol{\mu}$ instead of $\boldsymbol{v}$ although directly on $\boldsymbol{v}$ might be intuitively more understandable. We have Highlight 1.

Highlight 1: $\boldsymbol{v}$ suffers from two kinds of distributional uncertainties:

1) deviations imposed on mean and covariance [see (21), (22), and (23)];
2) deviations existing as outliers [see (24) and (25)].

However, the first one does not imply the second, and vice versa. They independently discredit the nominal Gaussian distribution of $\boldsymbol{v}$. □

Consequently, the dual problem (9) or (16) is equivalent to

$$\max_{\mathbb{P}(\boldsymbol{x})\in\mathcal{F}_{\boldsymbol{x}},\,\mathbb{P}(\boldsymbol{v})\in\mathcal{F}_{\boldsymbol{v}},\,\mathbb{P}(\mu)\in\mathcal{F}_{\mu}} \boldsymbol{P}, \tag{26}$$

where $\boldsymbol{P}$ is defined in (17); $\mathcal{F}_{\boldsymbol{x}}$, $\mathcal{F}_{\boldsymbol{v}}$, and $\mathcal{F}_{\mu}$ can be any types of possible ambiguity sets available above.

Theorem 2: Consider the dual problem (26). The following statements are true.

1) Reformulations for $\mathcal{F}_{\boldsymbol{x}}$.
   a) In (18), $\boldsymbol{c}_{\boldsymbol{x}}=\bar{\boldsymbol{x}}$ always holds so that $\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}}\|\bar{\mathbb{P}}_{\boldsymbol{x}})=\frac{1}{2}[\mathrm{Tr}[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}-\boldsymbol{I}]-\ln\det(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}})]$.
   b) In (19), $\boldsymbol{c}_{\boldsymbol{x}}=\bar{\boldsymbol{x}}$ always holds so that $\mathrm{W}(\mathbb{P}_{\boldsymbol{x}},\bar{\mathbb{P}}_{\boldsymbol{x}})=\sqrt{\mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}}+\boldsymbol{M}-2(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}]}$.
   c) In (20), $\boldsymbol{c}_{\boldsymbol{x}}=\bar{\boldsymbol{x}}$ always holds so that $\boldsymbol{\Sigma}_{\boldsymbol{x}}\preceq\theta_{2,\boldsymbol{x}}\boldsymbol{M}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}}\succeq\theta_{1,\boldsymbol{x}}\boldsymbol{M}$.
2) Reformulations for $\mathcal{F}_{\boldsymbol{v}}$.
   a) In (21), $\boldsymbol{c}_{\boldsymbol{v}}=\boldsymbol{0}$ always holds so that $\mathrm{KL}(\mathbb{P}_{\boldsymbol{v}}\|\bar{\mathbb{P}}_{\boldsymbol{v}})=\frac{1}{2}[\mathrm{Tr}[\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}-\boldsymbol{I}]-\ln\det(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}})]$.
   b) In (22), $\boldsymbol{c}_{\boldsymbol{v}}=\boldsymbol{0}$ always holds so that $\mathrm{W}(\mathbb{P}_{\boldsymbol{v}},\bar{\mathbb{P}}_{\boldsymbol{v}})=\sqrt{\mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{v}}+\boldsymbol{R}-2(\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}})^{\frac{1}{2}}]}$.
   c) In (23), $\boldsymbol{c}_{\boldsymbol{v}}=\boldsymbol{0}$ always holds so that $\boldsymbol{\Sigma}_{\boldsymbol{v}}\preceq\theta_{2,\boldsymbol{v}}\boldsymbol{R}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}\succeq\theta_{1,\boldsymbol{v}}\boldsymbol{R}$.
3) The dual problem (26) is equivalent to

$$\max_{\mathbb{P}(\boldsymbol{x})\in\mathcal{F}_{\boldsymbol{x}}}\max_{\mathbb{P}(\boldsymbol{v})\in\mathcal{F}_{\boldsymbol{v}}}\max_{\mathbb{P}(\mu)\in\mathcal{F}_{\mu}} \boldsymbol{P}. \tag{27}$$

The order of the three maximizations does not matter.

Proof: The estimation error covariance $\boldsymbol{P}$ in (26) does not depend on $\boldsymbol{c}_{\boldsymbol{x}}$ and $\boldsymbol{c}_{\boldsymbol{v}}$. In order to maximize $\boldsymbol{P}$, the larger the feasible sets of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, the better. This leads to $\boldsymbol{c}_{\boldsymbol{x}}=\bar{\boldsymbol{x}}$ and $\boldsymbol{c}_{\boldsymbol{v}}=\boldsymbol{0}$. Namely, the distributional uncertainty budgets $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{v}}$ are completely assigned to describe deviations of covariances of $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{P}_{\boldsymbol{v}}$, respectively, regardless of $\boldsymbol{c}_{\boldsymbol{x}}$ or $\boldsymbol{c}_{\boldsymbol{v}}$. This proves the first two claims 1) and 2). The claim 3) is standard in the optimization community. □

Since $\boldsymbol{s}=\boldsymbol{y}-\boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}}=\boldsymbol{H}(\boldsymbol{x}-\boldsymbol{c}_{\boldsymbol{x}})+\boldsymbol{v}$, and $\boldsymbol{x}$ and $\boldsymbol{v}$ are Gaussian and independent, the nominal value of $\boldsymbol{S}$ can be obtained as $\boldsymbol{S}=\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^T+\boldsymbol{\Sigma}_{\boldsymbol{v}}$. Let $i_\mu\in\mathbb{R}_+$ denote the Fisher information of $p(\mu)$; $i_\mu:=\mathbb{E}[-\frac{d^2}{d\mu^2}\ln p(\mu)]\geq 0$. Comparing with (17), $\boldsymbol{P}$ in (27) can be written as

$$\boldsymbol{P}=\boldsymbol{\Sigma}_{\boldsymbol{x}}-\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^T+\boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\cdot i_\mu.$$

In view of the first two claims 1) and 2) in Theorem 2, we identify that $\mathcal{F}_{\boldsymbol{x}}$ is parameterized by $\boldsymbol{\Sigma}_{\boldsymbol{x}}\in\mathbb{S}_+^n$ and $\mathcal{F}_{\boldsymbol{v}}$ is parameterized by $\boldsymbol{\Sigma}_{\boldsymbol{v}}\in\mathbb{S}_{++}^m$ (due to non-singularity of $\boldsymbol{S}$). Hence, (27) can be equivalently given as

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}}\max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}}\max_{i_\mu}\boldsymbol{P} \tag{28}$$

where the feasible sets of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ are defined by $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{v}}$, respectively. As a result, we can solve the reformulated dual problem (28) independently and sequentially, i.e., solving the innermost first and the outermost last.

The following two lemmas solve the innermost sub-problem over i_μ.

Lemma 1: The functional optimization over the ϵ-contamination ambiguity set

$$\min_{p(\mu)}\mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right]$$

$$s.t.\quad\begin{cases}\mu\sim\mathbb{P}_\mu\\ p(\mu)=\frac{d\mathbb{P}_\mu}{d\mu}\\ \mathbb{P}_\mu(\mu\in\mathbb{R})=1\\ \sup_\mu\|\mathbb{P}_\mu-\Phi(\cdot)\|\leq\epsilon\\ \mathbb{P}_\mu=(1-\epsilon)\Phi+\epsilon\mathbb{H}\\ \mathbb{H}(\mu)=1-\mathbb{H}(-\mu)\end{cases}$$

is solved by the following least-favorable distribution

$$p(\mu)=\begin{cases}(1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{K\mu+\frac{1}{2}K^2}, & \mu\leq -K\\ (1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\mu^2}, & |\mu|\leq K\\ (1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{-K\mu+\frac{1}{2}K^2}, & \mu\geq K,\end{cases} \tag{29}$$

where $K\in\mathbb{R}_+$ is implicitly defined by ϵ; $\int_{-K}^{K}p(\mu)dt+\frac{2p(K)}{K}=1$. Furthermore, $\min\mathbb{E}[-\frac{d^2}{d\mu^2}\ln p(\mu)]=(1-\epsilon)[1-2\Phi(-K)]$.

Proof: See Appendix D. □

Lemma 2: Given $0\leq\epsilon\lesssim 0.0303$, the functional optimization over the ϵ-normal ambiguity set

$$\min_{p(\mu)}\mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right]$$

$$s.t.\quad\begin{cases}\mu\sim\mathbb{P}_\mu\\ p(\mu)=\frac{d\mathbb{P}_\mu}{d\mu}\\ \mathbb{P}_\mu(\mu\in\mathbb{R})=1\\ \sup_\mu\|\mathbb{P}_\mu-\Phi(\cdot)\|\leq\epsilon\\ \mathbb{P}_\mu(\mu)=1-\mathbb{P}_\mu(-\mu)\end{cases}$$

is solved by the following least-favorable distribution

$$p(\mu)=\begin{cases}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}a^2}\cdot\cos^{-2}(\frac{1}{2}ca)\cdot\cos^2(\frac{1}{2}c\mu), & 0\leq\mu\leq a\\ \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\mu^2}, & a\leq\mu\leq b\\ \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}b^2}\cdot e^{-b\mu+b^2}, & \mu\geq b\end{cases} \tag{30}$$

and $p(\mu)=p(-\mu)$, where a, b, and c are implicitly defined by ϵ as

1) $c\tan(\frac{1}{2}ca)=a\ \ (0\leq ca<\pi)$,
2) $\int_0^a p(\mu)d\mu=\int_0^a d\Phi(\mu)-\epsilon$,
3) $\int_b^\infty p(\mu)d\mu=\int_b^\infty d\Phi(\mu)+\epsilon$.

Furthermore, $\min\mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right]=\frac{c^2a}{\cos^2(\frac{1}{2}ca)}p(a)+2\Phi(b)-2\Phi(a)$.

Proof: See Appendix E. □

Lemma 1 reveals that the least-favorable distribution under the ϵ-contamination distributional uncertainty is Gaussian in the middle (i.e., when $|\mu|\leq K$) and is Laplacian in the tails (i.e.,

when $\mu \geq K$ and $\mu \leq -K$), while Lemma 2 reveals that the least-favorable distribution under the ϵ-normal distributional uncertainty is $\cos^2(\cdot)$-like in the middle (i.e., when $-a \leq \mu \leq a$), is Gaussian in the transitions (i.e., when $a \leq \mu \leq b$ and $-b \leq \mu \leq -a$), and is Laplacian in the tails (i.e., when $\mu \geq b$ and $\mu \leq -b$). The Laplacian (a.k.a. exponential) tails (i.e., heavy tails) explain outliers in measurements. We call them least-favorable distributions because they have smallest Fisher information (i.e., largest asymptotic variance to estimate the mean) among distributions in $\mathcal{F}_\mu$. Although both are theoretically sound, we usually prefer the results in Lemma 1 because they coincide well with our intuitions from practice that the main part of measurements are normally distributed and only a small part of them are outliers. However, the results in Lemma 2 become suitable when quantization noises are non-negligible (e.g., when low-bit sampler is adopted), because quantization noise is close, but not equal, to zero.

Remark 2: In Lemma 2, we require that $\epsilon \lesssim 0.0303$ (n.b., for three real numbers, $x \lesssim z$ means that $x \leq y$ and $y \approx z$). This is a necessary condition to obtain the least-favorable distribution in (30). Otherwise, the least-favorable distribution is of a different form; see [39, p. 85 ff.]. Usually, we do not prefer the solution when $\epsilon \gtrsim 0.0303$ because the associated M-estimator has significantly larger asymptotic variance; see [39, Exhibit 4.6]. Theoretically, only when the true proportion of outliers is approximately smaller than 0.0303 can we use the solution in Lemma 2. However, in practice, the solution in Lemma 2 might not be sensitive to the true proportion of outliers: no matter what the true proportion of outliers (of course, as long as less than 0.5) in the true measurements, keeping $\epsilon \equiv 0.0303$ in our algorithm might not cause disasters. This observation is also true for the solution in Lemma 1. This point will be illustrated in the experiments in Subsection VII-D. □

After solving the innermost sub-problem of (28), we then study the outer sub-problems.

Theorem 3: The dual problem (28) is equivalent to

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}, \tag{31}$$

where $i_\mu^{\min} := \min i_\mu := \min \mathbb{E}[-\frac{d^2}{d\mu^2} \ln p(\mu)]$ is a constant defined in Lemma 1 or Lemma 2, whichever is adopted. Besides, $0 \leq i_\mu^{\min} \leq 1$.

Proof: Note that $\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0}$ because $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathbb{S}_+^n$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} \in \mathbb{S}_{++}^m$. Hence, the non-negative and minimal i_μ maximizes $\boldsymbol{P}$. In addition, since the standard Gaussian is contained in the ϵ-contamination set and the ϵ-normal set, $i_\mu^{\min}$ is upper bounded by the Fisher information of the standard Gaussian which is one. □

Since we have three alternatives for $\mathcal{F}_{\boldsymbol{x}}$, three for $\mathcal{F}_{\boldsymbol{v}}$, and two for $\mathcal{F}_\mu$, in principle, we need to solve the dual problem (26) eighteen times. As demonstrations and without loss of generality, we suppose $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{v}}$ have the same type of distributional uncertainty and study the distributionally robust Bayesian estimation (DRBE) under Wasserstein ambiguity and moment-based ambiguity, respectively.

Under Wasserstein ambiguities of $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{v}}$, the dual problem (31) can be explicitly written as

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}, \tag{32}$$

subject to

$$\begin{cases} \sqrt{\mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2\left(\boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \leq \theta_{\boldsymbol{x}} \\ \sqrt{\mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2\left(\boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{R}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \leq \theta_{\boldsymbol{v}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}. \end{cases} \tag{33}$$

This problem is difficult to solve as: 1) the objective is nonlinear, 2) the feasible set (33) is non-convex because the function $\sqrt{\cdot}$ is concave and the constraint $\sqrt{\mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}]} \leq \theta_{\boldsymbol{x}}$ is non-convex, so is the constraint for $\boldsymbol{v}$. However, we can still reformulate it into a linear semi-definite program (SDP) using some algebraic tricks. Solving a linear SDP is basic, although still challenging, in the optimization community.

Theorem 4: Suppose $\boldsymbol{R} \succ \boldsymbol{0}$. The dual problem (32) subject to (33) is solvable and can be reformulated as a linear SDP

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}, \boldsymbol{V}_{\boldsymbol{x}}, \boldsymbol{V}_{\boldsymbol{v}}, \boldsymbol{U}} \boldsymbol{\Sigma}_{\boldsymbol{x}} - i_\mu^{\min} \cdot \boldsymbol{U}, \tag{34}$$

subject to

$$\begin{cases} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T \\ \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix} \succeq \boldsymbol{0} \\ \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2\boldsymbol{V}_{\boldsymbol{x}}\right] \leq \theta_{\boldsymbol{x}}^2 \\ \begin{bmatrix} \boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{x}} \\ \boldsymbol{V}_{\boldsymbol{x}} & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0} \\ \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2\boldsymbol{V}_{\boldsymbol{v}}\right] \leq \theta_{\boldsymbol{v}}^2 \\ \begin{bmatrix} \boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{R}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{v}} \\ \boldsymbol{V}_{\boldsymbol{v}} & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{x}} \succeq \boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{v}} \succeq \boldsymbol{0}, \boldsymbol{U} \succeq \boldsymbol{0}. \end{cases} \tag{35}$$

Proof: See Appendix F. □

Under moment-based ambiguities of $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{v}}$, the dual problem (31) can be explicitly written as

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}, \tag{36}$$

subject to

$$\begin{cases} \boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \preceq \theta_{2,\boldsymbol{v}} \boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succeq \theta_{1,\boldsymbol{v}} \boldsymbol{R} \succ \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}. \end{cases} \tag{37}$$

This problem is relatively easier to solve than (32) because the feasible set (37) consists of linear constraints, implying convexity and compactness. Note that $\boldsymbol{R} \succ \boldsymbol{0}$ indicates $\boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}$. Therefore, it is solvable (i.e., the optimal solutions exist and are finite).

Theorem 5: The dual problem (36) subject to (37) is analytically solved by $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \theta_{2,\boldsymbol{x}} \boldsymbol{M}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \theta_{2,\boldsymbol{v}} \boldsymbol{R}$.

Proof: See Appendix G. □

It is also possible to jointly use the Wasserstein metric and the moment-based set, e.g., the Wasserstein metric for $\mathcal{F}_{\boldsymbol{x}}$ and the moment-based set for $\mathcal{F}_{\boldsymbol{v}}$. The derivations are straightforward and we do not cover the details.

As we can see, the dual problem under the moment-based distributional uncertainties admits attractive closed-form solutions which indicates high computational efficiency, especially for large scale estimation problems when n and m are (extremely) large. As for the problem under the Wasserstein metric, it requires solving a SDP which, although linear and solvable, is still computationally challenging. From the viewpoint of modelling, using the Wasserstein metric (33) [which is equivalent to (54) in Appendix F] or the moment-based set (37) just means that the shapes of the feasible sets are different. Since both (54) and (37) are convex and compact, for every $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ in (54), there exists $\theta_1 \in \mathbb{R}_+, \theta_2 \in \mathbb{R}_+$ for (37) such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ are contained in (37). Conversely, for every $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ in (37), there exists $\theta \in \mathbb{R}_+$ for (54) such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ are contained in (54). Therefore, in practice, we are not entangled in which type of ambiguity set we should choose. We use the one under which the problem is easy to solve. It is this reason that we do not study the problem under the KL divergence ambiguity in this article. Because nonlinear functions, i.e., $\ln(\cdot)$, $\det(\cdot)$, in (18) and (21) render the dual problem being a general nonlinear SDP (without linear reformulations) and difficult to solve. However, it is still convex and therefore solvable, since the constraints $\frac{1}{2}[\mathrm{Tr}[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{I}] - \ln\det(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}})] \leq \theta_{\boldsymbol{x}}$ and $\frac{1}{2}[\mathrm{Tr}[\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}} - \boldsymbol{I}] - \ln\det(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}})] \leq \theta_{\boldsymbol{v}}$ are convex. The convexity of the constraints is straightforward to show as: 1) $\mathrm{Tr}[\cdot]$ is linear and convex; 2) both $\ln(\cdot)$ and $\det(\cdot)$ are concave; 3) $\ln(\cdot)$ is monotonically increasing.

The theorem below summarizes the solution to the dual problem (9).

Theorem 6: Suppose the nominal distribution of $\boldsymbol{x}$ is $\bar{\mathbb{P}}_{\boldsymbol{x}} = \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and of $\boldsymbol{v}$ is $\bar{\mathbb{P}}_{\boldsymbol{v}} = \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{R}), \boldsymbol{R} \succ \boldsymbol{0}$. With Gaussianity assumptions for elements in the ambiguity sets $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{v}}$, the dual problem (9) is solved by

1) Optimal Estimator.

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}^* \boldsymbol{H}^T \boldsymbol{S}^{*-1/2} \cdot \boldsymbol{\psi}[\boldsymbol{S}^{*-1/2}(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}})], \quad (38)$$

where $\boldsymbol{S}^* := \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^* \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}}^*$, $\boldsymbol{\psi}(\mu)$ is entry-wise identical and for each entry

$$\psi(\mu) = \begin{cases} -K, & \mu \leq -K \\ \mu, & |\mu| \leq K \\ K, & \mu \geq K, \end{cases} \quad (39)$$

if the ϵ-contamination ambiguity set is used, or

$$\psi(\mu) = -\psi(-\mu) = \begin{cases} c\tan(\frac{1}{2}c\mu), & 0 \leq \mu \leq a \\ \mu, & a \leq \mu \leq b \\ b, & \mu \geq b, \end{cases} \quad (40)$$

if the ϵ-normal ambiguity set is used; $\boldsymbol{\Sigma}_{\boldsymbol{x}}^*$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}^*$ are the optimal solution of (34) if the Wasserstein metric is used, or of (36) if the moment-based set is used.

2) Worst-Case Estimation Error Covariance.

$$\boldsymbol{P}^* = \boldsymbol{\Sigma}_{\boldsymbol{x}}^* - \boldsymbol{\Sigma}_{\boldsymbol{x}}^* \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^* \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}}^*)^{-1} \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^* \cdot i_\mu^{\min}, \quad (41)$$

where

$$i_\mu^{\min} = (1-\epsilon)[1 - 2\Phi(-K)] \quad (42)$$

if the ϵ-contamination ambiguity set is used, or

$$i_\mu^{\min} = \frac{c^2 a}{\cos^2(\frac{1}{2}ca)} p(a) + 2\Phi(b) - 2\Phi(a) \quad (43)$$

if the ϵ-normal ambiguity set is used. For parameters K, a, b, and c, see Lemmas 1 and 2.

3) Least-Favorable Distributions.
   i) $\mathbb{P}_{\boldsymbol{x}}^* = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}^*, \boldsymbol{\Sigma}_{\boldsymbol{x}}^*)$, where $\boldsymbol{c}_{\boldsymbol{x}}^* = \bar{\boldsymbol{x}}$.
   ii) $\mathbb{P}_\mu^*$ is defined in (29) if the ϵ-contamination ambiguity set is used, or in (30) if the ϵ-normal ambiguity set is used.
   iii) $\mathbb{P}_{\boldsymbol{v}}^*$ is determined by the convolution of $\mathbb{P}_{\boldsymbol{\mu}}^*$ and $\mathbb{P}_{\boldsymbol{x}}^*$ through $\boldsymbol{v}^* = \boldsymbol{S}^{*\frac{1}{2}}\boldsymbol{\mu}^* - \boldsymbol{H}(\boldsymbol{x}^* - \bar{\boldsymbol{x}})$, where $\boldsymbol{S}^* := \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^* \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}}^*$. $\boldsymbol{v}^*$ denotes the random vector associated with $\mathbb{P}_{\boldsymbol{v}}^*$. Notations keep similar to $\boldsymbol{\mu}^*$ and $\boldsymbol{x}^*$.

Proof: See Appendix H. □

At last, we solve the primal distributionally robust Bayesian estimation problem (7).

Theorem 7: Under Gaussianity assumptions for nominal distributions of $\boldsymbol{x}$ and $\boldsymbol{v}$, the distributionally robust Bayesian estimation problem (7) admits the min-max property, i.e., the strong duality holds,

$$\min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{y}}} \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}) = \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})\in\mathcal{F}''} \min_{\boldsymbol{\phi}(\cdot)\in\mathcal{H}'_{\boldsymbol{y}}} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}),$$

where $\boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}) := \mathbb{E}_{\mathbb{P}(\cdot,\cdot)}[\boldsymbol{x} - \boldsymbol{\phi}(\boldsymbol{y})][\boldsymbol{x} - \boldsymbol{\phi}(\boldsymbol{y})]^T$. Hence, the solutions to the dual problem (9) also solve the primal problem (7).

Proof: See Appendix I. □

So far we have solved the distributionally robust Bayesian estimation problem subject to parameter uncertainties and measurement outliers. As a closing note, we mention that if we were sure that there are no outliers in measurements, we would have another modelling trick to address the distributionally robust Bayesian estimation problem. The theorem below is an outlier-free supplement to Theorem 6.

Theorem 8: If there are no outliers in measurements, we can directly model $\mathbb{P}_{\boldsymbol{x},\boldsymbol{y}}$ (or equivalently $\mathbb{P}_{\boldsymbol{x},\boldsymbol{v}}$) as a joint Gaussian distribution. In this special case, the ambiguity set admits $\Delta(\mathbb{P}_{\boldsymbol{x},\boldsymbol{y}}, \bar{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}) \leq \theta$, parameterized by just one scalar. $\Delta(\cdot,\cdot)$ can be any possible statistical metric or divergence (e.g., Wasserstein metric, KL divergence, moment-based set). If $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian, the optimal estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e., $\mathbb{E}(\boldsymbol{x}|\boldsymbol{y})$, has an affine form. Suppose the worst-case distribution is

$$\mathbb{P}_{\boldsymbol{x},\boldsymbol{y}}^* = \mathcal{N}_{n+m}\left(\begin{bmatrix} \boldsymbol{c}_{\boldsymbol{x}}^* \\ \boldsymbol{c}_{\boldsymbol{y}}^* \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^* & \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}}^* \\ \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{x}}^* & \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{y}}^* \end{bmatrix}\right).$$

We have the distributionally robust estimate as $\hat{\boldsymbol{x}} = \boldsymbol{c}_{\boldsymbol{x}}^* + \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}}^* \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{y}}^{*-1}(\boldsymbol{y} - \boldsymbol{c}_{\boldsymbol{y}}^*)$ and the worst-case estimation error covariance as $\boldsymbol{P}^* = \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^* - \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}}^* \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{y}}^{*-1} \boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{x}}^*$. Note that $\mathbb{P}_{\boldsymbol{x},\boldsymbol{y}}^*$ can be obtained in analogy to Theorem 4 if the Wasserstein metric is used, or to Theorem 5 if the moment-based set is used.

Proof: This special case has been discussed in [29]. □

When outliers exist in measurements, we can no longer assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ (or equivalently $\boldsymbol{x}$ and $\boldsymbol{v}$) are jointly Gaussian.

We have to separately discuss the ambiguity sets of $\mathbb{P}_{\boldsymbol{x}}$, $\mathbb{P}_{\boldsymbol{v}}$, $\mathbb{P}_{\mu}$, respectively. Even when there are no outliers in measurements, separately designing uncertainty sets for $\boldsymbol{x}$ and $\boldsymbol{y}$ (or equivalently $\boldsymbol{x}$ and $\boldsymbol{v}$) offers us more flexibility if we have different uncertain levels towards them, because jointly modelling admits the same uncertain levels.

V. Distributionally Robust State Estimation

With the results of distributionally robust Bayesian estimation developed in Section IV, this section solves the state estimation problem (6) at time k. We just need to identify the nominal conditional prior distribution of the state given the past measurements, i.e., $\bar{\mathbb{P}}_{\boldsymbol{x}_k|\boldsymbol{Y}_{k-1}}$. In our Gaussian approximation framework, $\bar{\mathbb{P}}_{\boldsymbol{x}_k|\boldsymbol{Y}_{k-1}}$ is Gaussian.

By (3), the nominal conditional prior distribution of the state $\boldsymbol{x}_k$ given the last state $\boldsymbol{x}_{k-1}$ is

$$\bar{\mathbb{P}}_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}} = \mathcal{N}_n\left(\boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1},\ \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}\boldsymbol{G}_{k-1}^T\right).$$

At time $k-1$, suppose the distributionally robust posterior state estimate is $\mathbb{E}(\boldsymbol{x}_{k-1}|\boldsymbol{Y}_{k-1}) := \hat{\boldsymbol{x}}_{k-1|k-1}$ and the associated estimation error covariance is $\boldsymbol{P}^*_{k-1|k-1}$; the conditional distribution of $\boldsymbol{x}_{k-1}$ given $\boldsymbol{Y}_{k-1}$ is $\mathbb{P}_{\boldsymbol{x}_{k-1}|\boldsymbol{Y}_{k-1}} = \mathcal{N}_n(\hat{\boldsymbol{x}}_{k-1|k-1},\ \boldsymbol{P}^*_{k-1|k-1})$. Therefore, the nominal conditional prior distribution of the state $\boldsymbol{x}_k$ given $\boldsymbol{Y}_{k-1}$ is

$$\bar{\mathbb{P}}_{\boldsymbol{x}_k|\boldsymbol{Y}_{k-1}} = \int_{\mathbb{R}^n} \bar{\mathbb{P}}(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1})\, \mathbb{P}(d\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}), \tag{44}$$

i.e.,

$$\bar{\mathbb{P}}(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}) \sim \mathcal{N}_n\left(\hat{\boldsymbol{x}}_{k|k-1}, \boldsymbol{M}_{k|k-1}\right), \tag{45}$$

where

$$\hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}_{k-1}\hat{\boldsymbol{x}}_{k-1|k-1} \tag{46}$$

and

$$\boldsymbol{M}_{k|k-1} = \boldsymbol{F}_{k-1}\boldsymbol{P}^*_{k-1|k-1}\boldsymbol{F}_{k-1}^T + \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}\boldsymbol{G}_{k-1}^T. \tag{47}$$

The nominal distribution of the measurement noise $\boldsymbol{v}_k$ is $\bar{\mathbb{P}}_{\boldsymbol{v}_k|\boldsymbol{Y}_{k-1}} = \bar{\mathbb{P}}_{\boldsymbol{v}_k} = \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{R}_k)$ because we can readily verify that $\boldsymbol{v}_k$ is independent of $\boldsymbol{Y}_{k-1}$.

Now it is sufficient to invoke the results in Theorem 6 to obtain the distributionally robust state estimate $\hat{\boldsymbol{x}}_{k|k}$ at time k given $\boldsymbol{y}_k$.

Theorem 9: Suppose the radii of the ambiguity sets are $\epsilon \geq 0$, $\theta_{\boldsymbol{x},k} \geq 0$, $\theta_{2,\boldsymbol{x},k} \geq 1 \geq \theta_{1,\boldsymbol{x},k} \geq 0$, $\theta_{\boldsymbol{v},k} \geq 0$, $\theta_{2,\boldsymbol{v},k} \geq 1 \geq \theta_{1,\boldsymbol{v},k} \geq 0$. At time k, with the nominal Gaussian prior conditional distribution of the state $\bar{\mathbb{P}}_{\boldsymbol{x}_k|\boldsymbol{Y}_{k-1}} \sim \mathcal{N}_n(\hat{\boldsymbol{x}}_{k|k-1}, \boldsymbol{M}_{k|k-1})$ and the nominal Gaussian distribution of the measurement noise $\bar{\mathbb{P}}_{\boldsymbol{v}_k} = \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{R}_k)$, the distributionally robust state estimate $\hat{\boldsymbol{x}}_{k|k}$ given $\boldsymbol{y}_k$ is as follows.

1) Optimal Estimator.

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}_k^T\boldsymbol{S}_k^{*-1/2} \cdot \boldsymbol{\psi}[\boldsymbol{S}_k^{*-1/2}\boldsymbol{s}_k], \tag{48}$$

where $\boldsymbol{s}_k := \boldsymbol{y}_k - \boldsymbol{H}_k\hat{\boldsymbol{x}}_{k|k-1}$, $\hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}_{k-1}\hat{\boldsymbol{x}}_{k-1|k-1}$, and $\boldsymbol{S}^*_k := \boldsymbol{H}_k\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}_k^T + \boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$; $\boldsymbol{\psi}(\cdot)$, $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}$, and $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$ are defined in Theorem 6.

2) Worst-Case Estimation Error Covariance.

$$\boldsymbol{P}^*_{k|k} = \boldsymbol{\Sigma}^*_{\boldsymbol{x},k} - \boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}_k^T\boldsymbol{S}_k^{*-1}\boldsymbol{H}_k\boldsymbol{\Sigma}^*_{\boldsymbol{x},k} \cdot i_\mu^{\min}, \tag{49}$$

where $i_\mu^{\min}$ is defined in Theorem 6.

Algorithm 1: Distributionally Robust Estimator.

Definition: $\hat{\boldsymbol{x}}_{k|k}$ as the distributionally robust state estimate; $\boldsymbol{P}^*_{k|k}$ as the worst-case state estimation error covariance.
Initialize: $\hat{\boldsymbol{x}}_{0|0}$, $\boldsymbol{P}^*_{0|0}$, ϵ, all involved θ as instructed in Theorem 9 (i.e., $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{v}}$ if we use the Wasserstein ambiguity sets, and $\theta_{2,\boldsymbol{x}}$ and $\theta_{2,\boldsymbol{v}}$ if we use the moment-based ambiguity sets).
Remark: According to Theorem 5, $\theta_{1,\boldsymbol{x}}$ and $\theta_{1,\boldsymbol{v}}$ are irrelevant to this algorithm, and therefore, not initialized.
Input: $\boldsymbol{y}_k$, $k = 1, 2, 3, \ldots$
while true **do**
  // Time-Update Step, i.e., Prior Estimation
    Use (46) and (47) to obtain $\hat{\boldsymbol{x}}_{k|k-1}$ and $\boldsymbol{M}_{k|k-1}$
  // Obtain the Nominal Distributions
    Use (45) to obtain $\bar{\mathbb{P}}_{\boldsymbol{x}_k|\boldsymbol{Y}_{k-1}}$
    $\bar{\mathbb{P}}_{\boldsymbol{v}_k} \leftarrow \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{R}_k)$
  // Obtain the Worst-Case Scenario
    Use (42) or (43) to obtain $i_\mu^{\min}$
    Use (34) or (36) to obtain $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}$ and $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$
  // Measurement-Update Step, i.e., Posterior Estimation
    Use (48) and (49) to obtain $\hat{\boldsymbol{x}}_{k|k}$ and $\boldsymbol{P}^*_{k|k}$
    // Next Time Step
    $k \leftarrow k + 1$
end while
Output $\hat{\boldsymbol{x}}_{k|k}$

Proof: Compare with Theorem 6. □

The distributionally robust estimator to the linear Markov system (3) is summarized in Algorithm 1.

The theorem below reveals relations among the proposed distributionally robust estimator and the existing estimators.

Theorem 10: Concerning the distributionally robust state estimator in Algorithm 1, the follows are true.

1) If we set $\epsilon = 0$, $\theta_{\boldsymbol{x}} = \theta_{\boldsymbol{v}} = 0$, $\theta_{1,\boldsymbol{x}} = \theta_{2,\boldsymbol{x}} = 1$, $\theta_{1,\boldsymbol{v}} = \theta_{2,\boldsymbol{v}} = 1$, we obtain the canonical Kalman filter.
2) Under moment-based ambiguities, if we set $\epsilon = 0$, $\theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}}$, we obtain the fading Kalman filter [4], [5].
3) The Student's t Kalman filter in [31, Eq. (13)] amounts to a distributionally robust filter because it is a fading Kalman filter whose fading factor is adaptively changeable.
4) Under moment-based ambiguities, if we set $\epsilon = 0$, $\theta_{1,\boldsymbol{x}} = \theta_{2,\boldsymbol{x}} = 1$, we obtain the robust Kalman filter in [61, Eq. (32)] that has an adaptive $\theta_{2,\boldsymbol{v}}$.
5) Under ϵ-contamination ambiguity, if we set $\theta_{\boldsymbol{x}} = \theta_{\boldsymbol{v}} = 0$, $\theta_{1,\boldsymbol{x}} = \theta_{2,\boldsymbol{x}} = 1$, $\theta_{1,\boldsymbol{v}} = \theta_{2,\boldsymbol{v}} = 1$, we obtain the M-estimation-based Kalman filter [34, Thm. 3].
6) When there are no outliers and the special case discussed in Theorem 8 is considered, if we use the Wasserstein metric, we obtain the Wasserstein Kalman filter [28].
7) When there are no outliers and the special case discussed in Theorem 8 is considered, if we use the KL divergence, we obtain the relative-entropy Kalman filter [26].
8) When there are no outliers in measurements and the special case discussed in Theorem 8 is considered, if we use the τ-divergence, we obtain the τ-divergence Kalman filter [27].
9) The relative-entropy Kalman filter and the τ-divergence Kalman filter are risk-sensitive Kalman filters [26], [27].

Proof: In the case 1), all the ambiguity sets only contain nominal distributions. Hence, we have $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k} = \boldsymbol{M}_{k|k-1}$, $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k} = \boldsymbol{R}_k$, $\psi(\mu) = \mu$, and $i^{\min}_{\mu} = 1$, leading to the canonical Kalman filter. In the case 2), if we assume $\theta = \theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}}$, we have $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k} = \theta \boldsymbol{M}_{k|k-1}$, $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k} = \theta \boldsymbol{R}_k$, $\psi(\mu) = \mu$, and $i^{\min}_{\mu} = 1$, leading to $\boldsymbol{P}^*_{k|k} = \theta \cdot \boldsymbol{P}_{k|k}$ where $\boldsymbol{P}_{k|k} := \boldsymbol{M}_{k|k-1} - \boldsymbol{M}_{k|k-1}\boldsymbol{H}_k^T(\boldsymbol{H}_k\boldsymbol{M}_{k|k-1}\boldsymbol{H}_k^T + \boldsymbol{R}_k)^{-1}\boldsymbol{H}_k\boldsymbol{M}_{k|k-1}$. By comparing with [5], we obtain the fading Kalman filter. For other cases, compare with the given references. □

VI. Comparisons With Existing Frameworks

A. Frameworks Addressing Parameter Uncertainties

As a typical framework, we first review the unknown-input filters below. In the unknown-input filters (e.g., [19]), the following linear system is studied

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{\Gamma}_{k-1}\boldsymbol{d}_{k-1} + \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1}, \\ \boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k, \end{cases} \tag{50}$$

where $\boldsymbol{d}_{k-1} \in \mathbb{R}^q$ is the unknown input describing the parameter uncertainties (e.g., $\boldsymbol{F}_{k-1}$ and/or the mean of $\boldsymbol{w}_{k-1}$ are no longer exact). In this case, the parameter uncertainties are limited within the range space of $\boldsymbol{\Gamma}_{k-1}$. Therefore, to guarantee satisfactory performances, the structure and entries of $\boldsymbol{\Gamma}_{k-1}$ need to be carefully designed. A wrong choice of $\boldsymbol{\Gamma}_{k-1}$ would, on the contrary, lead to catastrophic results.

The next solutions are the robust state estimation methods. Robust state estimation methods aim to make the filters insensitive to parameter uncertainties. They robustify the filters through minimizing the worst-case mean square error matrix (a.k.a., upper bound of estimation error covariance [62], [63]), although the uncertainties are described, structured, parameterized, and bounded in different ways. Among existing literature, two classic frameworks are remarkable. In [16], the following perturbed linear system is studied

$$\begin{cases} \boldsymbol{x}_k = (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\boldsymbol{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\boldsymbol{w}_{k-1}, \\ \boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k, \end{cases} \tag{51}$$

where $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ are parameter perturbations. They are assumed to be additive and imposed on the nominal system matrices $\boldsymbol{F}_{k-1}$ and $\boldsymbol{G}_{k-1}$, respectively. Besides, $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ are assumed to hold the following structure

$$\left[\delta\boldsymbol{F}_{k-1}\ \delta\boldsymbol{G}_{k-1}\right] = \boldsymbol{M}_{k-1}\Delta_{k-1}\left[\boldsymbol{E}_{f,k-1}\ \boldsymbol{E}_{g,k-1}\right], \tag{52}$$

where Δ_{k-1} is an arbitrary contraction operator whose operator norm is less than one. $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, and $\boldsymbol{E}_{g,k-1}$ are structural matrices to be carefully designed based on our reliable knowledge about the uncertainties. Therefore, intuitively, the parameters' uncertainty space is expressed by limiting the matrix norms of the system perturbation matrices $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$. In [24], another model about $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ is considered. In this case, $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ are assumed to be linear combinations of random variables, i.e.,

$$\begin{cases} \delta\boldsymbol{F}_{k-1} = \sum_{i=1}^{l} \boldsymbol{F}_{i,k-1} \cdot \zeta_{i,k-1} \\ \delta\boldsymbol{G}_{k-1} = \sum_{i=1}^{l} \boldsymbol{G}_{i,k-1} \cdot \zeta_{i,k-1}, \end{cases} \tag{53}$$

where $\zeta_{i,k-1}$ are random variables with assumed-known statistics; l, $\boldsymbol{F}_{i,k-1}$, and $\boldsymbol{G}_{i,k-1}$ are assumed to be exactly known. In this case, the second moment of the state vector are confined in a convex and compact polytope and the worst-case estimation is obtained over it. However, the final state estimation formulation is a SDP which is challenging to solve.

As we can see, in order to describe the parameters' uncertainties, we must have reliable information about them so that we can choose proper structures and entries of $\boldsymbol{\Gamma}_{k-1}$, $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, $\boldsymbol{E}_{g,k-1}$, $\boldsymbol{F}_{i,k-1}$, and $\boldsymbol{G}_{i,k-1}$, and determine exact statistics of $\zeta_{i,k-1}$. For some specific problems, this is possible, while for general ones, it is hard. This emphasizes the advantages of the proposed distributionally robust state estimation framework, which does not require the structural information of parameters' uncertainties.

B. Frameworks Addressing Measurement Outliers

When we unexpectedly see outliers in a nominal outlier-free population, we usually have two philosophies. The first one is that we no longer believe the nominal population is outlier-free. Instead, we take into account the outliers directly in modelling and correct the nominal distribution into an outlier-aware one. Typical solutions include: 1) direct modelling, e.g., t-distribution, Laplacian distribution; 2) indirect modelling, e.g., Bayesian methods (e.g., if the variance of a Gaussian distribution follows an inverse Gamma distribution, then the samples from this variance-variant Gaussian distribution would follow a t-distribution). The second one is that we still believe the population is outlier-free and treat seen outliers as aggressors to be cleared/modified. Typical solutions are reported, in particular, by Frequentists, e.g., the jackknife method.

The two philosophies can also be understood by leveraging the influence curve (a.k.a. influence function; see Appendix C) [38], [39], [64]. Two kinds of influence curves are well-studied:

1) infinite-rejection-point influence curves, including all the monotonic influence curves (e.g., Huber's [43]) and some re-descending influence curves that have infinite rejection points (e.g., maximum-correntropy-criterion [37], [65]).
2) finite-rejection-point influence curves, including some re-descending influence curves that have finite rejection points (e.g., Hampel's [38], [39], Tukey's Biweight [39], Andrew's Sine [39], IGG [66]).

When we use infinite-rejection-point influence curves, we implicitly accept outliers to be unstudied samples and correct the nominal distribution to be heavy-tailed. For example, the influence curve of an M-estimator at a t-distribution is a kind of re-descending influence curve but it has infinite rejection-point [33, Fig. 1]. Contrarily, when we adopt finite-rejection-point influence curves, we actually admit finite support of the nominal distribution and any sample outside of this support would be treated as intruders and trashed.

Most of the existing state estimation frameworks under measurement outliers belong to one of the two philosophies mentioned above. Note that in Bayesians, different from Frequentists, influence curves are imposed on innovation vectors (i.e., difference between true measurement and predicted measurement; cf. Theorem 6) rather than directly on measurement vectors; see, e.g., [34]. Below lists and discusses some typically existing outlier-insensitive state estimation frameworks.

The earliest outlier-treatment method is the Gaussian-sum filter [3], [30], which uses heavy-tailed distributions for measurements, and the non-Gaussian heavy-tailed distributions are approximated by Gaussian sums. The demerit of this method is that it is computationally intensive and, thus, inefficient.

A remedy methodology to the Gaussian-sum filter is typically the t-distribution Kalman filter [31], [32], [67], which no longer uses a Gaussian sum to approximate the non-Gaussian measurement noise. Instead, it directly uses heavy-tailed non-Gaussian distributions such as the t-distribution, which explicitly explain the outliers. An indirect modelling trick is the Bayesian framework that assumes the noise statistics matrix (i.e., $\boldsymbol{R}$) is not exact and follows an inverse Wishart distribution so that the measurements $\boldsymbol{y}$ from the linear observation $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{v}$ would follow a multivariate t-distribution, which implicitly accounts for outliers [33].

Another remedy methodology is directly working on designing proper influence functions [37], [38], which is also known as the weighted-least-square M-estimation-based Kalman filter [35], [36]. For details, see Appendix C. In this category, the solutions for $\psi(\cdot)$ defined in Theorem 6 and Theorem 9 are particularly popular. Other possible influence functions are the maximum-correntropy-criterion (MCC) [65], IGG [66], Hampel's [38], [39], Tukey's Biweight [39], Andrew's Sine [39], etc. However, note that they are derived from other motivations and might no longer have clear perspectives of distributional robustness.

VII. Experiments

In this section, we compare the state estimation performances of the existing filters and our newly proposed filter. All the source data and codes are available online at GitHub: https://github.com/Spratm-Asleaf/DRSE-Outlier. Interested readers can reproduce and/or verify the claims in this article via changing the parameters or codes by themselves. Additional experiments are placed in the online supplementary materials.

We continue studying the classical instance discussed in [16], [26], [28], i.e.,

$$\boldsymbol{F}_k^{real} = \begin{bmatrix} 0.9802 & 0.0196 + \alpha \cdot \Delta_k \\ 0 & 0.9802 \end{bmatrix},$$

$$\boldsymbol{G}_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{H}_k = \begin{bmatrix} 1 & -1 \end{bmatrix},$$

$$\boldsymbol{Q}_k = \begin{bmatrix} 1.9608 & 0.0195 \\ 0.0195 & 1.9605 \end{bmatrix}, \boldsymbol{R}_k = \begin{bmatrix} 1 \end{bmatrix},$$

where the random scalar $\Delta_k \in \mathcal{U} := [-1, 1]$ denotes the real perturbations imposed on the system and $\mathcal{U}$ defines its support; α is a multiplicative coefficient. In this state estimation problem, the nominal system matrix is known as

$$\boldsymbol{F}_k = \begin{bmatrix} 0.9802 & 0.0196 \\ 0 & 0.9802 \end{bmatrix}.$$

Besides, we randomly add outliers for 5% measurements (i.e., we accordingly set $\epsilon = 0.05$ in the proposed method).

A. Candidate Filters

We implement the following filters to compare.

1) **TMKF**: the canonical Kalman filter with the true model. In the simulation we know the underlying true model $\boldsymbol{F}_k^{real}$ and the outlier-free true measurements. Therefore, this method theoretically gives the best estimate of state in the sense of minimum estimation error covariance;
2) **KF**: the canonical Kalman filter (with the nominal model $\boldsymbol{F}_k$);
3) **HKF**: the outlier-insensitive Kalman filter based on the Huber's influence function [34], [37];
4) **τ-KF**: the τ-divergence Kalman filter [27];
5) **WKF**: the Wasserstein Kalman filter [28];
6) **MKF**: The moment-based distributionally robust state estimator (see Theorem 9). We choose moment-based ambiguity sets because under them the problem is easier to solve (than that under Wasserstein ambiguity sets).

B. Parameters Setting

Algorithm 1 requires to initialize the parameters ϵ and θs. Note that when ϵ is specified, K in (39), and a, b, and c in (40) will be uniquely determined; see Lemmas 1 and 2 and their proofs. Besides, if we use the Wasserstein ambiguity sets, we need to initialize $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{v}}$ [see (33)]. If we use the moment-based ambiguity sets, we need to initialize $\theta_{2,\boldsymbol{x}}$ and $\theta_{2,\boldsymbol{v}}$ [see (37)] (n.b., Algorithm 1 is irrelevant to $\theta_{1,\boldsymbol{x}}$ and $\theta_{1,\boldsymbol{v}}$).

In all methods, we set the initial state estimate as $\hat{\boldsymbol{x}}_{0|0} = [0, 0]^T$ and its corresponding estimation error covariance as $\boldsymbol{P}^*_{0|0} := \mathrm{diag}\{1, 1\}$, where $\mathrm{diag}\{\cdot\}$ denotes a diagonal matrix [16], [26], [28]. All parameters of each filter are directly taken from the original paper or tuned to perform (nearly) best for the studied instance when Δ_k randomly changes and $\alpha = 1$.

In the Huber-based outlier-insensitive Kalman filter, we use $K = 1.4$ [see (39)], because when ϵ is fixed to 0.05, K has to be 1.4 [cf. (29)]. In the τ-divergence Kalman filter [27], we set $\tau = 0$ (i.e., the τ-divergence filter specifies the Kullback-Leibler filter [26]), and the radius of the ambiguity set as 1.5×10^{-4}. In the Wasserstein Kalman filter [28], the radius of the ambiguity set is set to 0.1. In the moment-based distributionally robust filter, we set $\theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}} = 1.02$, and $K = 1.4$. Namely, the influence function in (39) is used.

Suppose each simulation episode runs $T = 1000$ discrete-time steps. The overall estimation error of each episode is measured by the rooted mean square error (RMSE) as

$$\sqrt{\frac{1}{T}\sum_{k=1}^{T}[(x_{1,k} - \hat{x}_{1,k})^2 + (x_{2,k} - \hat{x}_{2,k})^2]},$$

where $x_{1,k}$ (resp. $x_{2,k}$) is the first (resp. second) component of the state vector $\boldsymbol{x}_k$ and $\hat{x}_{1,k}$ (resp. $\hat{x}_{2,k}$) denotes its estimate.

C. Results

Results are obtained by a laptop with 8 G RAM and Intel(R) Core(TM) i7-8850H CPU @ 2.60 GHz. We conduct the following three experiments, respectively. First, let Δ_k randomly take its value according to the uniform distribution from its support $\mathcal{U}$ at each step k, and let $\alpha = 1$. However, in this simulation, we do not add outliers in the measurements. The results are shown in Table I. Second, let $\alpha = 0$ (i.e., there are no parameter uncertainties). Nevertheless, we add outliers for 5% measurements. The

TABLE I
RESULTS WHEN $\alpha = 1$ BUT NO OUTLIERS

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**3.25**	1.41e-5	τ-KF [27]	**9.90**	25.25e-5
KF	14.52	7.51e-6	WKF [28]	9.95	132.24e-5
HKF [34]	14.74	1.22e-5	MKF[Ours]	**9.91**	1.23e-5

Avg Time: Average Execution Time at each time step (seconds); **1e-5**: 1×10^{-5}; **Note:** TMKF gives theoretically optimal solution.

TABLE II
RESULTS WHEN $\alpha = 0$ AND ONLY OUTLIERS

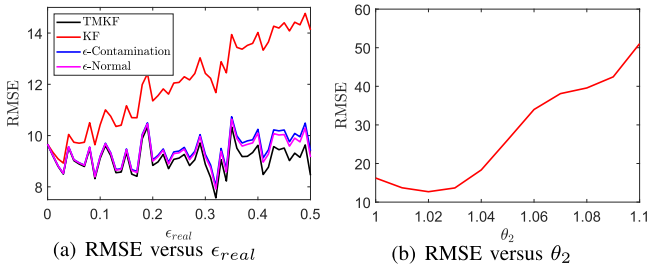
Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**7.64**	1.36e-5	τ-KF [27]	19.41	26.43e-5
KF	16.14	7.82e-6	WKF [28]	16.56	125.55e-5
HKF [34]	**7.70**	1.39e-5	MKF[Ours]	**8.19**	1.20e-5

See Table I for table notes.

TABLE III
RESULTS WHEN $\alpha = 1$ AND ALSO OUTLIERS

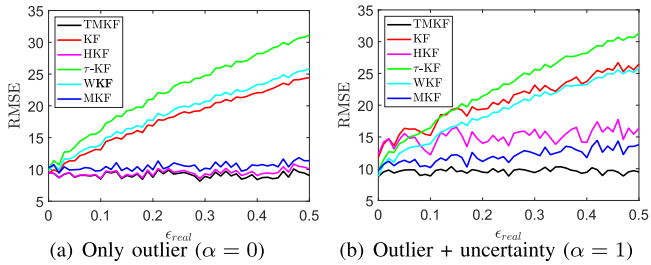
Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**3.23**	1.40e-5	τ-KF [27]	22.15	25.76e-5
KF	21.04	7.31e-6	WKF [28]	16.94	126.52e-5
HKF [34]	16.14	1.26e-5	MKF[Ours]	**11.72**	1.16e-5

See Table I for table notes.



(a) RMSE versus ϵ_{real} (b) RMSE versus θ_2

Fig. 1. Sensitivity results over ϵ_{real} and θ_2.



(a) Only outlier ($\alpha = 0$) (b) Outlier + uncertainty ($\alpha = 1$)

Fig. 2. Breakdown test against ϵ_{real} with and without model uncertainty.

results are shown in Table II. Third, both parameter uncertainties and measurement outliers are considered as above. The results are shown in Table III.

From Tables I, II, and III, the following observations are outlined. When there only exist parameter uncertainties, the τ-divergence Kalman filter, the Wasserstein Kalman filter, and the proposed moment-based distributionally robust state estimator are relatively robust, while the Huber-based outlier-insensitive Kalman filter is not. In addition, the proposed moment-based distributionally robust state estimator is preferable since it is computationally efficient. When there only exist measurement outliers, the Huber-based outlier-insensitive Kalman filter is roughly optimal as expected. However, the τ-divergence Kalman filter and the Wasserstein Kalman filter perform badly, implying that they are not robust against measurement outliers. When both parameter uncertainties and measurement outliers exist, the proposed moment-based distributionally robust state estimator works better than other candidate filters; i.e., it is robust against both parameter uncertainties and measurement outliers. In Tables I and III, the performances of the proposed method are far away from those of the TMKF because a relatively large uncertainty coefficient α is used (i.e., the true system model is far away from the nominal one). When α is set to be small, the difference will reduce (cf. Table II). This reminds us that the robust filters are just remedial, but not once-for-all, solutions. In practice, continuing efforts need to be put on improving the accuracy of the nominal model, unless the model accuracy cannot be refined or robust solutions are satisfactory.

D. Sensitivity Analysis

In reality, it is hard to know the exact values of the true proportion of outliers (i.e., ϵ), and the true uncertainty level of the nominal model (i.e., $\theta_{\boldsymbol{x}}$, $\theta_{\boldsymbol{v}}$, $\theta_{2,\boldsymbol{x}}$, and $\theta_{2,\boldsymbol{v}}$). They cannot be learned to be optimal either because for a real system, the true state is unknown (i.e., training data set is unavailable). Hence, we need to investigate whether the proposed algorithm is sensitive to parameters ϵ and θs, and explore the prior knowledge of tuning them for a real problem. Without loss of generality, we continue using the instance discussed above, where the moment-based ambiguity sets are adopted. As before, we set $\theta_{2,\boldsymbol{x}}$ and $\theta_{2,\boldsymbol{v}}$ to be the same, and $\theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}} := \theta_2$.

First, we let $\alpha = 0$ (i.e., no model uncertainty) and only study the sensitivity against the true proportion of outliers. For the case that we use the influence function in (39), we arbitrarily set $\epsilon = 0.01$ so that $K = 2$; for the case that we use the influence function in (40), we let $\epsilon = 0.03$ so that $a = 1.3496$, $b = 1.3496$, and $c = 1.2316$. Then, we let the real proportion of outliers ϵ_{real} change from 0 to 0.5. We have the results in Fig. 1(a). It shows that the proposed method is not sensitive to ϵ_{real}. Thus, it is safe in practice to keep the values of ϵ, K, a, b, and c recommended above regardless of ϵ_{real}. (Other values are also viable; readers can validate this claim using the shared source codes themselves.) Besides, we show the breakdown properties of all the candidate filters. The results are shown in Fig. 2. We see that the HKF is better than the MKF when there are no model uncertainties [cf. Fig. 2(a)], whereas the HKF is worse than the MKF when there exist model uncertainties [cf. Fig. 2(b)]. This is because the MKF is the robustified version of the HKF against model uncertainties (n.b., the MKF reduces to the HKF when $\theta_2 := 0$). Therefore, the price of the robustness in uncertain conditions (when $\alpha \neq 0$) is sacrificing the optimality in perfect conditions (when $\alpha = 0$).

Second, we fix $\epsilon_{real} = 0.05$ and study the sensitivity against the true degree of the model uncertainty. We let $\alpha = 1$, and θ_2 change from 1 to 1.1. We have the results in Fig. 1(b). It shows that the performance of the proposed method depends heavily on the value of θ_2. If θ_2 is too small, the algorithm has no sufficient robustness against the uncertainty. Contrarily, if θ_2 is too large, the algorithm is too conservative to obtain a good performance as

well. Therefore, one should carefully (and pragmatically) tune this parameter to achieve good performances for their specific real problems.

VIII. Conclusion

This article proposes the distributionally robust state estimation method that can account for both parameter uncertainties and measurement outliers. It offers a new perspective to understand the robust state estimation problem under parameter uncertainties and measurement outliers and generalizes several classic methods into a unified framework. It uses only a few scalars to describe parameter uncertainties and measurement outliers and does not require structural information of uncertainties, especially useful when we have limited trust towards the nominal model and scarce knowledge about the uncertainties. Experiments show that the proposed method under moment-based ambiguity sets outperforms existing methods, which is not hard to expect because none of them is designed to simultaneously address both parameter uncertainties and measurement outliers. Although the method might be insensitive to the true proportion of outliers (i.e., the value of ϵ used in the algorithm does not significantly matter), it is sensitive to the true uncertainty level of the nominal model (i.e., the values of θs used in the algorithm significantly matter). Practitioners have to carefully try appropriate θs for their specific problems (n.b., θs cannot be learned because the true state is unavailable). At last, two closing remarks need to be outlined. 1) Robust filters are just remedial solutions. Reducing modelling uncertainties is always important. Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known). 2) The robustness under uncertain conditions comes with the cost of sacrificing the optimality under perfect conditions.

Appendix A
Proof of Theorem 1

The conditional mean $\hat{\boldsymbol{x}} = [p(\boldsymbol{y})]^{-1}\int \boldsymbol{x}p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} = [p(\boldsymbol{y})]^{-1}\int(\boldsymbol{x}-\bar{\boldsymbol{x}})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} + \bar{\boldsymbol{x}} = [p(\boldsymbol{y})]^{-1}\boldsymbol{M}\int p_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\boldsymbol{M}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})p(\boldsymbol{x})d\boldsymbol{x} + \bar{\boldsymbol{x}}$. Due to the prior distribution of $\boldsymbol{x}$ is Gaussian, $-\boldsymbol{M}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})p(\boldsymbol{x}) = \frac{dp(\boldsymbol{x})}{d\boldsymbol{x}}$, giving $\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} - \boldsymbol{M}[p(\boldsymbol{y})]^{-1}\int p_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\frac{dp(\boldsymbol{x})}{d\boldsymbol{x}}d\boldsymbol{x}$. By partial integration, we have $\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \boldsymbol{M}[p(\boldsymbol{y})]^{-1}\int \frac{\partial p_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})}{\partial \boldsymbol{x}} p(\boldsymbol{x})d\boldsymbol{x} = \bar{\boldsymbol{x}} - \boldsymbol{M}\boldsymbol{H}^T[p(\boldsymbol{y})]^{-1}\int \frac{\partial p_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})}{\partial \boldsymbol{y}} p(\boldsymbol{x})d\boldsymbol{x} = \bar{\boldsymbol{x}} - \boldsymbol{M}\boldsymbol{H}^T[p(\boldsymbol{y})]^{-1}\int \frac{\partial p(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}}d\boldsymbol{x} = \bar{\boldsymbol{x}} - \boldsymbol{M}\boldsymbol{H}^T[p(\boldsymbol{y})]^{-1}\frac{dp(\boldsymbol{y})}{d\boldsymbol{y}} = \bar{\boldsymbol{x}} + \boldsymbol{M}\boldsymbol{H}^T[-\frac{d\ln p(\boldsymbol{y})}{d\boldsymbol{y}}]$. Furthermore, the conditional covariance of the estimation error is $\boldsymbol{P}_{\boldsymbol{x}|\boldsymbol{y}} = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(\hat{\boldsymbol{x}}-\boldsymbol{x})(\cdots)^T = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(\bar{\boldsymbol{x}}-\boldsymbol{x})(\cdots)^T - \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(\hat{\boldsymbol{x}}-\bar{\boldsymbol{x}})(\cdots)^T = \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^T[-\frac{d\ln p(\boldsymbol{y})}{d\boldsymbol{y}}][\cdots]^T\boldsymbol{H}\boldsymbol{M}$. Since $\boldsymbol{\mu} = \boldsymbol{S}^{-1/2}(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})$, we have $p_{\boldsymbol{\mu}}(\boldsymbol{\mu}) = p_{\boldsymbol{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu}+\boldsymbol{H}\bar{\boldsymbol{x}})\cdot\det[\frac{d(\boldsymbol{S}^{1/2}\boldsymbol{\mu}+\boldsymbol{H}\bar{\boldsymbol{x}})}{d\boldsymbol{\mu}}] = p_{\boldsymbol{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu}+\boldsymbol{H}\bar{\boldsymbol{x}})\cdot\det(\boldsymbol{S}^{1/2})$. As a result, $-\frac{d\ln p(\boldsymbol{\mu})}{d\boldsymbol{\mu}} = -\frac{d\ln p_{\boldsymbol{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu}+\boldsymbol{H}\bar{\boldsymbol{x}})}{d\boldsymbol{\mu}} = -\boldsymbol{S}^{1/2}\frac{d\ln p_{\boldsymbol{y}}(\boldsymbol{y})}{d\boldsymbol{y}}$, implying $-\frac{d\ln p_{\boldsymbol{y}}(\boldsymbol{y})}{d\boldsymbol{y}} = \boldsymbol{S}^{-1/2}[-\frac{d\ln p(\boldsymbol{\mu})}{d\boldsymbol{\mu}}]$. Similarly, we can show that $[-\frac{d\ln p_{\boldsymbol{y}}(\boldsymbol{y})}{d\boldsymbol{y}}][\cdots]^T = \boldsymbol{S}^{-1/2}[-\frac{d\ln p(\boldsymbol{\mu})}{d\boldsymbol{\mu}}][\cdots]^T\boldsymbol{S}^{-1/2}$. Combining the derivations above, we finish the proof. □

Appendix B
Proof of Corollary 1

By noting that $\mathbb{E}\{[p(\boldsymbol{\mu})]^{-1}\frac{d^2p(\boldsymbol{\mu})}{d\boldsymbol{\mu}\boldsymbol{\mu}^T}\} = \int[p(\boldsymbol{\mu})]^{-1}\frac{d^2p(\boldsymbol{\mu})}{d\boldsymbol{\mu}\boldsymbol{\mu}^T}p(\boldsymbol{\mu})d\boldsymbol{\mu} = \frac{d^2\int p(\boldsymbol{\mu})d\boldsymbol{\mu}}{d\boldsymbol{\mu}\boldsymbol{\mu}^T} = \frac{d^2 1}{d\boldsymbol{\mu}\boldsymbol{\mu}^T} = \boldsymbol{0}$, we have $\mathbb{E}[-\frac{d^2\ln p(\boldsymbol{\mu})}{d\boldsymbol{\mu}\boldsymbol{\mu}^T}] = -\mathbb{E}\{[p(\boldsymbol{\mu})]^{-1}\frac{d^2p(\boldsymbol{\mu})}{d\boldsymbol{\mu}\boldsymbol{\mu}^T}\} + \mathbb{E}\{[p(\boldsymbol{\mu})]^{-2}[-\frac{dp(\boldsymbol{\mu})}{d\boldsymbol{\mu}}][\cdots]^T\} = \mathbb{E}[-\frac{d\ln p(\boldsymbol{\mu})}{d\boldsymbol{\mu}}][\cdots]^T$. □

Appendix C
Some Statistical Concepts

Suppose the density of interest $p(\mu;\theta)$ is parameterized by unknown mean θ. For mean estimation (a.k.a. location estimation) problems, in general, $p(\mu;\theta) := p(\mu-\theta)$; recall, e.g., the Gaussian distribution. Strictly speaking, the score function is defined with respect to the unknown parameter θ as $\frac{d}{d\theta}\ln p(\mu;\theta)$. Since $\frac{d}{d\theta}\ln p(\mu;\theta) = -\frac{d}{d\mu}\ln p(\mu;\theta)$, in statistics, some authors also directly define the score function with respect to μ as $-\frac{d}{d\mu}\ln p(\mu;\theta)$. As a result, the Fisher information has two equivalent definitions as well: $\mathbb{E}[-\frac{d^2}{d\theta^2}\ln p(\mu;\theta)]$ and $\mathbb{E}[-\frac{d^2}{d\mu^2}\ln p(\mu;\theta)]$.

In statistics, the three concepts, score function, influence function, and weight function, are closely related but different. Score function is well-known in maximum likelihood estimation, influence function in general (outlier-) robust statistics [38], [39, Chap. 3], and weight function in (outlier-) robust linear regression [35], [36], [68], [39, Chap. 7]. Influence function is a property of an estimator designed for a distribution, while score function is that of the distribution itself. However, in M-estimation, influence function is just a multiple of score function and the constant multiplier is the Fisher information associated with the distribution. Let $T_\theta(\mathbb{P}_\mu)$ be the M-estimator of the mean of the distribution $\mathbb{P}_\mu$ whose density is $p(\mu)$. Supposing a score function is given by $\psi(\mu) := -\frac{d}{d\mu}\ln p(\mu)$, the influence function $IF(\mu)$ equals to [38]

$$IF(\mu) := \lim_{\epsilon\downarrow 0}\frac{T_\theta[(1-\epsilon)\mathbb{P}_\mu + \epsilon\Delta_\mu] - T_\theta[\mathbb{P}_\mu]}{\epsilon} = \frac{\psi(\mu)}{-\int\psi'(\mu)p(\mu)d\mu},$$

where Δ_μ is a point mass distribution concentrated at μ, $\psi'(\cdot)$ is the derivative of $\psi(\cdot)$, and the denominator is the Fisher information. In particular, if the Fisher information of the distribution is unit (e.g., standard Gaussian), the score function coincides with the influence function. For this reason, in M-estimation contexts, practitioners first derive score function and then equate it to influence function because a score function is mathematically easier to obtain. On the other hand, the weight function in (outlier-) robust linear regression is defined by $\psi(\mu)/\mu$. In statistical theory the three concepts are distinguished because they have different backgrounds, meanings, and definitions, but in signal processing practice we consider them to be equivalent (in the sense that one uniquely implies another) because they

have similar mathematical forms. With this implication in mind, it is not confusing that the score function $\psi(\cdot)$ shown in Theorem 6 and Theorem 9 is directly termed as "influence function" in literature such as [33], [34], [40], [69]. This is more intuitively understandable for signal processing practitioners because $\psi(\cdot)$ limits the "influence" that a (contaminated) measurement $\boldsymbol{y}_k$ may bring.

In M-estimation contexts, when we mention to design an influence function, we mean to design the score function $\psi(\cdot)$ [38]. Besides, when we design a weight function in robust linear regression contexts, we also uniquely obtain the corresponding score function in M-estimation counterpart [37]. The score function, in turn, implicitly determines the distribution for the studied population (which includes both ordinary points and outliers); $p(\mu) \propto \exp\left[-\int_{-\infty}^{\mu} \psi(\mu) d\mu\right]$ because $\psi(\mu) = -\frac{d}{d\mu} \ln p(\mu)$.

Appendix D
Proof of Lemma 1

See [43, pp. 80] for the solution of $p(\mu)$. As a result, $\min \mathbb{E}[-\frac{d^2}{d\mu^2} \ln p(\mu)] = \int_{-K}^{K} p(\mu) dt = (1-\epsilon) \int_{-K}^{K} d\Phi(t) = (1-\epsilon)[1 - 2\Phi(-K)]$. For any given ϵ, the value of K can be found in [43, Table I] or [39, Exhibit 4.3]. □

Appendix E
Proof of Lemma 2

See [43, pp. 91] for the solution of $p(\mu)$. As a result, $\min \mathbb{E}[-\frac{d^2}{d\mu^2} \ln p(\mu)] = 2 \times [\int_0^a \frac{1}{2} \frac{c^2}{\cos^2(\frac{1}{2}c\mu)} p(\mu) d\mu + \int_a^b p(\mu) d\mu] = 2[\frac{1}{2} \frac{c^2}{\cos^2(\frac{1}{2}ca)} p(a) \int_0^a d\mu + \int_a^b d\Phi(\mu)]$. For any given $0 \leq \epsilon \lesssim 0.0303$, the values of a, b, and c can be found in [43, Table II] or [39, Exhibit 4.6]. □

Appendix F
Proof of Theorem 4

The squared constraint $\mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}] \leq \theta_{\boldsymbol{x}}^2$ is convex and compact, so is the squared constraint for $\boldsymbol{v}$ (as $\boldsymbol{R} \succ \boldsymbol{0}$) [53]. Therefore, the following equivalent feasible set is convex and also compact.

$$\begin{cases} \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2\left(\boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \theta_{\boldsymbol{x}}^2 \\ \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2\left(\boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{R}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \theta_{\boldsymbol{v}}^2 \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}. \end{cases} \tag{54}$$

Due to $\boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}$, the existence of the inverse in the objective function (32) is guaranteed. As the trace of the objective (32) is continuous, smooth (i.e., differentiable), and joint concave in terms of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, the dual problem (32) subject to (33) is solvable (i.e., the optimal solutions exist and are finite).

In order to simplify the objective function, let $\boldsymbol{U} \succeq \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0}$. By Schur complement, it is equivalent to require

$$\begin{bmatrix} \boldsymbol{U} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T \\ \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix} \succeq \boldsymbol{0}.$$

In order to simplify the constraints, let $\boldsymbol{V}_{\boldsymbol{x}} \preceq (\boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}$, i.e., $\boldsymbol{V}_{\boldsymbol{x}}^2 \preceq \boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}}$. By Schur complement, it is equivalent to require

$$\begin{bmatrix} \boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{x}} \\ \boldsymbol{V}_{\boldsymbol{x}} & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0}.$$

Likewise, let $\boldsymbol{V}_{\boldsymbol{v}} \preceq (\boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{R}^{\frac{1}{2}})^{\frac{1}{2}}$, i.e., $\boldsymbol{V}_{\boldsymbol{v}}^2 \preceq \boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{R}^{\frac{1}{2}}$. By Schur complement, it is equivalent to require

$$\begin{bmatrix} \boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{R}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{v}} \\ \boldsymbol{V}_{\boldsymbol{x}} & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0}.$$

Note that $\boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{M}^{\frac{1}{2}} \succeq \boldsymbol{0}$ and $\boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{v}} \boldsymbol{R}^{\frac{1}{2}} \succeq \boldsymbol{0}$. □

Appendix G
Proof of Theorem 5

In order to simplify the objective function, let $\boldsymbol{U} \preceq \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_{\mu}^{\min}$. By Schur complement, the dual problem (36) subject to (37) is equivalent to

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}, \boldsymbol{U}} \boldsymbol{U},$$

subject to

$$\begin{cases} \begin{bmatrix} (\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{U})/i_{\mu}^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T \\ \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix} \succeq \boldsymbol{0} \\ \boldsymbol{U} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \preceq \theta_{2,\boldsymbol{v}} \boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succeq \theta_{1,\boldsymbol{v}} \boldsymbol{R} \succ \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}. \end{cases}$$

Namely,

$$\begin{bmatrix} \boldsymbol{U}/i_{\mu}^{\min} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \preceq \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}}/i_{\mu}^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T \\ \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix}.$$

Since $\boldsymbol{I} \succ \boldsymbol{0}$ and $\boldsymbol{I}/i_{\mu}^{\min} - \boldsymbol{H}(\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \succeq \boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \succeq \boldsymbol{0}$, by Schur complement, we have

$$\frac{d \,\mathrm{Tr}\begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}}/i_{\mu}^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T \\ \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix}}{d \boldsymbol{\Sigma}_{\boldsymbol{x}}} = \begin{bmatrix} \boldsymbol{I}/i_{\mu}^{\min} & \boldsymbol{H} \\ \boldsymbol{H}^T & \boldsymbol{H}^T \boldsymbol{H} \end{bmatrix} \succeq \boldsymbol{0},$$

and

$$\frac{d \,\mathrm{Tr}\begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}}/i_{\mu}^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T \\ \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix}}{d \boldsymbol{\Sigma}_{\boldsymbol{v}}} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0},$$

implying the upper bound of $\begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}}/i_{\mu}^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T \\ \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix}$ is reached by the upper bounds of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$. Note that $\boldsymbol{H}(\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T$ is an idempotent matrix (a.k.a. projection matrix in linear regression) whose eigenvalues only contain zeros and ones (therefore, $\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \succeq \boldsymbol{0}$).

As a result,

$$\begin{bmatrix} \boldsymbol{U}/i_{\mu}^{\min} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \preceq \begin{bmatrix} \theta_{2,\boldsymbol{x}}\boldsymbol{M}/i_{\mu}^{\min} & \theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^T \\ \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M} & \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^T + \theta_{2,\boldsymbol{v}}\boldsymbol{R} \end{bmatrix},$$

giving

$$\begin{bmatrix} (\theta_{2,\boldsymbol{x}}\boldsymbol{M} - \boldsymbol{U})/i_{\mu}^{\min} & \theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^T \\ \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M} & \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^T + \theta_{2,\boldsymbol{v}}\boldsymbol{R} \end{bmatrix} \succeq \boldsymbol{0}.$$

Therefore, the upper bound of $\boldsymbol{U}$ is

$$\theta_{2,\boldsymbol{x}}\boldsymbol{M} - \theta_{2,\boldsymbol{x}}^2\boldsymbol{M}\boldsymbol{H}^T(\boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^T + \theta_{2,\boldsymbol{v}}\boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{M} \cdot i_{\mu}^{\min},$$

reached by $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \theta_{2,\boldsymbol{x}}\boldsymbol{M}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \theta_{2,\boldsymbol{v}}\boldsymbol{R}$. □

APPENDIX H
PROOF OF THEOREM 6

By noting that $\boldsymbol{\psi}(\cdot) := -\frac{d}{d\boldsymbol{\mu}} \ln p(\boldsymbol{\mu})$ and recalling (29) and (30) in the worst case, Eqs. (39) and (40) are immediate. For the worst-case distribution of $\boldsymbol{v}$, it is not simply $\mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}^*, \boldsymbol{\Sigma}_{\boldsymbol{v}}^*)$ where $\boldsymbol{c}_{\boldsymbol{v}}^* = \boldsymbol{0}$ because we have $\boldsymbol{v} = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{\mu} - \boldsymbol{H}(\boldsymbol{x} - \bar{\boldsymbol{x}})$. From Highlight 1, the distribution of $\boldsymbol{v}$ suffers from two types of deviations, i.e., outlier-related and outlier-unrelated. $\mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}^*, \boldsymbol{\Sigma}_{\boldsymbol{v}}^*)$ is just the worst-case distribution for the outlier-unrelated part. The integrated worst-case distribution of $\boldsymbol{v}$ is determined by the convolution of $\mathbb{P}_{\boldsymbol{\mu}}^*(\boldsymbol{\mu})$ and $\mathbb{P}_{\boldsymbol{x}}^*(\boldsymbol{x})$ through $\boldsymbol{v}^* = \boldsymbol{S}^{*\frac{1}{2}}\boldsymbol{\mu}^* - \boldsymbol{H}(\boldsymbol{x}^* - \bar{\boldsymbol{x}})$. It is non-trivial to explicitly compute this convolution. However, fortunately, we do not need to pursue its exact expression (or numerical value). The other statements are straightforward from Lemmas 1, 2 and Theorems 1, 2, 3, 4, 5. □

APPENDIX I
PROOF OF THEOREM 7

The weak duality admits

$$\max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}''} \min_{\boldsymbol{\phi}(\cdot) \in \mathcal{H}_{\boldsymbol{y}}'} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}) \preceq \min_{\boldsymbol{\phi}(\cdot) \in \mathcal{H}_{\boldsymbol{y}}'} \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}''} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}).$$

Supposing the estimator $\boldsymbol{\phi}^*$ and the worst case distribution $\mathbb{P}^*$ solve the dual problem which are available from Theorem 6, we have $\boldsymbol{V}(\boldsymbol{\phi}^*, \mathbb{P}^*) \preceq \min_{\boldsymbol{\phi}(\cdot) \in \mathcal{H}_{\boldsymbol{y}}'} \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}''} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P})$. On the other hand,

$$\min_{\boldsymbol{\phi}(\cdot) \in \mathcal{H}_{\boldsymbol{y}}'} \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}''} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}) \preceq \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}''} \boldsymbol{V}(\boldsymbol{\phi}^*, \mathbb{P}).$$

Since $\mathbb{P}^*$ maximizes the right hand side (see Theorem 6),

$$\min_{\boldsymbol{\phi}(\cdot) \in \mathcal{H}_{\boldsymbol{y}}'} \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}''} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}) \preceq \boldsymbol{V}(\boldsymbol{\phi}^*, \mathbb{P}^*).$$

As a result,

$$\min_{\boldsymbol{\phi}(\cdot) \in \mathcal{H}_{\boldsymbol{y}}'} \max_{\mathbb{P}(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}''} \boldsymbol{V}(\boldsymbol{\phi}, \mathbb{P}) = \boldsymbol{V}(\boldsymbol{\phi}^*, \mathbb{P}^*).$$

This shows the min-max property, i.e., strong duality, completing the proof. □

REFERENCES

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960, doi: 10.1115/1.3662552..

[2] D. Simon, *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[3] G. Chen, *Approximate Kalman Filtering*. Singapore: World Scientific Publishing Company, 1993.

[4] T. J. Tarn and J. Zaborszky, "A practical nondiverging filter," *AIAA J.*, vol. 8, no. 6, pp. 1127–1133, 1970.

[5] G. Gawrys and V. Vandelinde, "Divergence and the fading memory filter," in *Proc. IEEE Conf. Decis. Control Including 14th Symp. Adaptive Processes*, 1975, pp. 66–68.

[6] B. Hassibi, A. H. Sayed, and T. Kailath, "Linear estimation in Krein spaces. II. applications," *IEEE Trans. Autom. Control*, vol. 41, no. 1, pp. 34–49, Jan. 1996.

[7] Y. S. Shmaliy, F. Lehmann, S. Zhao, and C. K. Ahn, "Comparing robustness of the Kalman, H_∞, and UFIR filters," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3447–3458, Jul. 2018.

[8] J. Speyer, J. Deyst, and D. Jacobson, "Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria," *IEEE Trans. Autom. Control*, vol. AC-19, no. 4, pp. 358–366, Aug. 1974.

[9] D. Bertsekas and I. Rhodes, "Recursive state estimation for a set-membership description of uncertainty," *IEEE Trans. Autom. Control*, vol. AC-16, no. 2, pp. 117–128, Apr. 1971.

[10] X. Shen and L. Deng, "Game theory approach to discrete H_∞ filter design," *IEEE Trans. Signal Process.*, vol. 45, no. 4, pp. 1092–1095, Apr. 1997.

[11] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Trans. Autom. Control*, vol. AC-15, no. 2, pp. 175–184, Apr. 1970.

[12] A. Mohamed and K. Schwarz, "Adaptive Kalman filtering for INS/GPS," *J. Geodesy*, vol. 73, no. 4, pp. 193–203, 1999.

[13] Y. Huang, Y. Zhang, Z. Wu, N. Li, and J. Chambers, "A novel adaptive Kalman filter with inaccurate process and measurement noise covariance matrices," *IEEE Trans. Autom. Control*, vol. 63, no. 2, pp. 594–601, Feb. 2018.

[14] Y. Huang, Y. Zhang, P. Shi, and J. Chambers, "Variational adaptive Kalman filter with Gaussian-inverse-Wishart mixture distribution," *IEEE Trans. Autom. Control*, vol. 66, no. 4, pp. 1786–1793, Apr. 2021.

[15] I. R. Petersen and A. V. Savkin, *Robust Kalman Filtering for Signals and Systems With Large Uncertainties*. New York, NY, USA: Springer Science & Business Media, 1999.

[16] A. H. Sayed, "A framework for state-space estimation with uncertain models," *IEEE Trans. Autom. Control*, vol. 46, no. 7, pp. 998–1013, Jul. 2001.

[17] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, "Interacting multiple model methods in target tracking: A survey," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, no. 1, pp. 103–123, Jan. 1998.

[18] Y. Ma, S. Zhao, and B. Huang, "Multiple-model state estimation based on variational Bayesian inference," *IEEE Trans. Autom. Control*, vol. 64, no. 4, pp. 1679–1685, Apr. 2019.

[19] S. Gillijns and B. De Moor, "Unbiased minimum-variance input and state estimation for linear discrete-time systems," *Automatica*, vol. 43, no. 1, pp. 111–116, 2007.

[20] J. George, "A robust estimator for stochastic systems under unknown persistent excitation," *Automatica*, vol. 63, pp. 156–161, 2016.

[21] S. Z. Yong, M. Zhu, and E. Frazzoli, "A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems," *Automatica*, vol. 63, pp. 321–329, 2016.

[22] S. Wang, C. Li, and A. Lim, "Optimal joint estimation and identification theorem to linear Gaussian system with unknown inputs," *Signal Process.*, vol. 161, pp. 268–288, 2019.

[23] U. Shaked, L. Xie, and Y. C. Soh, "New approaches to robust minimum variance filter design," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2620–2629, Nov. 2001.

[24] F. Wang and V. Balakrishnan, "Robust Kalman filters for linear time-varying systems with stochastic parametric uncertainties," *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 803–813, Apr. 2002.

[25] W. Liu and P. Shi, "Convergence of optimal linear estimator with multiplicative and time-correlated additive measurement noises," *IEEE Trans. Autom. Control*, vol. 64, no. 5, pp. 2190–2197, May 2019.

[26] B. C. Levy and R. Nikoukhah, "Robust state space filtering under incremental model perturbations subject to a relative entropy tolerance," *IEEE Trans. Autom. Control*, vol. 58, no. 3, pp. 682–695, Mar. 2013.

[27] M. Zorzi, "Robust Kalman filtering under model perturbations," *IEEE Trans. Autom. Control*, vol. 62, no. 6, pp. 2902–2907, Jun. 2017.

[28] S. S. Abadeh, V. A. Nguyen, D. Kuhn, and P. M. M. Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8474–8483.

[29] S. Wang, Z. Wu, and A. Lim, "Robust state estimation for linear systems under distributional uncertainty," *IEEE Trans. Signal Process.*, vol. 69, pp. 5963–5978, Oct. 2021, doi: 10.1109/TSP.2021.3118540.
[30] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.
[31] Y. Huang, Y. Zhang, N. Li, Z. Wu, and J. A. Chambers, "A novel robust student's *t*-based Kalman filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 3, pp. 1545–1554, Jun. 2017.
[32] Y. Huang, Y. Zhang, Y. Zhao, and J. A. Chambers, "A novel robust Gaussian-student's *t* mixture distribution based Kalman filter," *IEEE Trans. Signal Process.*, vol. 67, no. 13, pp. 3606–3620, Jul. 2019.
[33] G. Agamennoni, J. I. Nieto, and E. M. Nebot, "Approximate inference in state-space models with heavy-tailed noise," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5024–5037, Oct. 2012.
[34] C. Masreliez and R. Martin, "Robust Bayesian estimation for the linear model and robustifying the Kalman filter," *IEEE Trans. Autom. Control*, vol. 22, no. 3, pp. 361–371, Jun. 1977.
[35] Z. M. Durovic and B. D. Kovacevic, "Robust estimation with unknown noise statistics," *IEEE Trans. Autom. Control*, vol. 44, no. 6, pp. 1292–1296, Jun. 1999.
[36] M. A. Gandhi and L. Mili, "Robust Kalman filter based on a generalized maximum-likelihood-type estimator," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2509–2520, May 2010.
[37] L. Chang and K. Li, "Unified form for the robust Gaussian information filtering based on M-estimate," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 412–416, Apr. 2017.
[38] F. R. Hampel, "The influence curve and its role in robust estimation," *J. Amer. Stat. Assoc.*, vol. 69, no. 346, pp. 383–393, 1974.
[39] P. J. Huber, *Robust Statistics*, 2nd ed. Hoboken, NJ, USA: Wiley, 2009.
[40] V. Stojanovic, S. He, and B. Zhang, "State and parameter joint estimation of linear stochastic systems in presence of faults and non-Gaussian noises," *Int. J. Robust Nonlinear Control*, vol. 30, no. 16, pp. 6683–6700, 2020.
[41] D. A. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*. New York, NY, USA: Wiley, 1954.
[42] H. Scarf, "A min max solution of an inventory problem," in *Studies in the Mathematical Theory of Inventory and Production*. Stanford, CA, USA: Stanford Univ. Press, 1958.
[43] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964.
[44] D. Bertsimas, M. Sim, and M. Zhang, "Adaptive distributionally robust optimization," *Manage. Sci.*, vol. 65, no. 2, pp. 604–618, 2019.
[45] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.
[46] M. Staib and S. Jegelka, "Distributionally robust optimization and generalization in kernel methods," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9134–9144.
[47] I. Yang, "A dynamic game approach to distributionally robust safety specifications for stochastic systems," *Automatica*, vol. 94, pp. 94–101, 2018.
[48] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," *INFORMS Tut. Operations Res.*, pp. 130–166, 2019, doi: 10.1287/educ.2019.0198.
[49] A. Ben-Tal, D. D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Manage. Sci.*, vol. 59, no. 2, pp. 341–357, 2013.
[50] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Res.*, vol. 58, no. 3, pp. 595–612, 2010.
[51] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.
[52] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ, USA: Prentice Hall, 2000.
[53] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization," *Math. Operations Res.*, 2021, doi: 10.1287/moor.2021.1176..
[54] B. Hassibi, A. H. Sayed, and T. Kailath, "Linear estimation in Krein spaces. I. Theory," *IEEE Trans. Autom. Control*, vol. 41, no. 1, pp. 18–33, Jan. 1996.
[55] X. Shen and P. K. Varshney, "Sensor selection based on generalized information gain for target tracking in large sensor networks," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 363–375, Jan. 2014.
[56] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Autom. Control*, vol. 54, no. 6, pp. 1254–1269, Jun. 2009.
[57] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proc. IEEE Adaptive Syst. Signal Process., Commun., Control Symp. (Cat. No. 00EX373)*, 2000, pp. 153–158.
[58] K. Kim and G. Shevlyakov, "Why gaussianity?," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 102–113, Mar. 2008.
[59] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3072–3081, Nov. 2001.
[60] D. Guo, Y. Wu, S. S. Shitz, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, Apr. 2011.
[61] Z. Li, Y. Yao, J. Wang, and J. Gao, "Application of improved robust Kalman filter in data fusion for PPP/INS tightly coupled positioning system," *Metrol. Meas. Syst.*, vol. 24, no. 2, pp. 289–301, 2017.
[62] L. Xie, Y. C. Soh, and C. E. DeSouza, "Robust Kalman filtering for uncertain discrete-time systems," *IEEE Trans. Autom. Control*, vol. 39, no. 6, pp. 1310–1314, Jun. 1994.
[63] Y. Liang, D. Zhou, L. Zhang, and Q. Pan, "Adaptive filtering for stochastic systems with generalized disturbance inputs," *IEEE Signal Process. Lett.*, vol. 15, pp. 645–648, Oct. 2008, doi: 10.1109/LSP.2008.2002707.
[64] F. Hampel, "Contributions to the theory of robust estimation," Ph.D. dissertation, Univ. California, Berkeley, Sep. 1968.
[65] B. Chen, X. Liu, H. Zhao, and J. C. Principe, "Maximum correntropy Kalman filter," *Automatica*, vol. 76, pp. 70–77, 2017.
[66] Y. Yuanxi, "Robust estimation for dependent observations," *Manuscripta Geodaetica*, vol. 19, no. 1, pp. 10–17, 1994.
[67] L. Sun, W. K. Ho, K. V. Ling, T. Chen, and J. Maciejowski, "Recursive maximum likelihood estimation with *t*-distribution noise model," *Automatica*, vol. 132, 2021, Art. no. 109789.
[68] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, Jul. 2012.
[69] S. Wang, Z. Wu, and A. Lim, "Denoising, outlier/dropout correction, and sensor selection in range-based positioning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021, Art. no. 1007613.

Shixiong Wang (Student Member, IEEE) received the B.Eng. degree in detection, guidance and control technology, and the M.Eng. degree in systems and control engineering from the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, in 2016 and 2018, respectively. He is currently working toward the Ph.D. degree with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore.

His research interests include statistics and optimization theories with applications in signal processing (especially optimal estimation theory) and control technology.

Zhi-Sheng Ye (Senior Member, IEEE) received the joint B.E. degree in material science and engineering and economics from Tsinghua University, Beijing, China, in 2008, and the Ph.D. degree in industrial and systems engineering from the National University of Singapore, Singapore, in 2012.

He is currently an Associate Professor in industrial engineering with the Department of Industrial Systems Engineering and Management, National University of Singapore. His research interests include reliability engineering, complex systems modeling, and industrial statistics.

Prof. Ye is an Associate Editor for the IEEE TRANSACTIONS ON RELIABILITY and the *IISE Transactions*.