Optimal Bayesian Kalman Filtering With Prior Update

Roozbeh Dehghannasiri, *Student Member, IEEE*, Mohammad Shahrokh Esfahani, *Member, IEEE*, Xiaoning Qian, *Senior Member, IEEE*, and Edward R. Dougherty, *Fellow, IEEE*

***Abstract*—In many practical filter design problems, the exact statistical information of the underlying random processes is not available. One robust filtering approach in these situations is to design an intrinsically Bayesian robust filter that provides optimal solution relative to the prior distribution governing the uncertainty class of all possible joint random process models. In this context, the intrinsically Bayesian robust Kalman filter has been recently introduced for the case that the second-order statistics of the observation and process noise in the state-space model are unknown. However, such a filter does not utilize the additional information embedded in the data being observed. In this paper, we derive the optimal Bayesian Kalman filter, which is optimal over posterior distribution obtained from incorporating data into the prior distribution. This filter has the same recursive structure as that of the classical Kalman filter, except that it is designed relative to the *posterior effective noise statistics*, which are found by employing the method of factor graphs through formulating the problem of computing the likelihood function as a message passing algorithm.**

***Index Terms*—Kalman filter, uncertain noise statistics, Bayesian robustness, posterior distribution, factor graphs.**

I. Introduction

DESIGNING an optimal filter requires the exact knowledge of the statistical model involving underlying random processes; however, in a wide variety of engineering applications, it is not possible to perfectly identify the model due to complexity, practical limitations, limited data, etc. Therefore, it is of interest to design a robust filter that behaves well relative to an uncertainty class of possible models compatible with partial prior knowledge. Designing robust filters has been an active area of research dating back to the late 1970s and early 1980s, examples being Wiener filtering with regards to uncertain spectra [1], [2], matched filtering [3], and Kalman filtering [4], [5]. These early works treat robust filtering from a minimax perspective, the aim being to find a filter with the best worst-case performance across the uncertainty class. The downside to the minimax approach is that it might be overly influenced by the extreme models in the uncertainty class. Later works on robust filtering take a Bayesian view for handling uncertainty where the aim is to find a filter with the best performance on average relative to the prior distribution [6]–[11].

The Kalman filter [12] has been widely used in engineering applications such as navigation, target tracking, and inference. This filter has a simple structure and can be implemented in the time domain; however, it is highly sensitive to the accuracy of the noise statistics [13]. Consequently, designing robust Kalman filters has drawn much attention. There are different approaches for robust Kalman filtering. Most involve adaptive Kalman filtering, where the aim is to simultaneously estimate noise statistics along with state estimation [14]–[17]. Generally, adaptive Kalman methods can be divided into Bayesian, maximum likelihood, correlation, and covariance matching methods [17]. The problem with adaptive Kalman filters is that they usually need a large number of observations to tune their parameters and achieve reliable performance. Other robust Kalman filtering approaches are minimax Kalman filters [4], [5], [18] and finite impulse response Kalman filters [19]–[21].

The authors have recently proposed an *intrinsically Bayesian robust* (IBR) Kalman filter that provides minimum average MSE relative to the prior distribution of the unknown noise covariances [9]. This filter has a recursive structure similar to that of the classical Kalman filter, except that it uses *effective* noise statistics and an *effective* Kalman gain matrix, which are extended versions of their classical counterparts when applied to the uncertainty class. The IBR theory for Kalman filtering utilizes the notions of Bayesian innovation process and Bayesian orthogonality principle [9]. The concept of innovation process has a long history in signal processing, rooted in the seminal works of Bode, Shannon, Zadeh, and Ragazzini in the 1950s and 1960s [22], [23]. Later Kailath used this approach to derive the Kalman filter recursive equations [24].

Although the IBR approach provides optimal filtering relative to the prior distribution, it does not utilize the new information provided by the observations. In this paper, we extend the IBR Kalman filter to the *optimal Bayesian Kalman filter* (OBKF), where the optimization is relative to the posterior distribution obtained after updating the prior distribution by utilizing data. To this end, we revisit the results for the IBR Kalman filter and state them relative to the posterior distribution. The proposed OBKF framework involves similar recursive equations as those of the classical Kalman filter with the *posterior effective noise statistics* in place of the ordinary noise statistics. Posterior effective noise statistics represent the posterior distribution of the noise

Manuscript received May 18, 2017; revised November 11, 2017; accepted December 20, 2017. Date of publication January 4, 2018; date of current version March 1, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. D. Robert Iskander. *(Corresponding author: Roozbeh Dehghannasiri.)*

R. Dehghannasiri, X. Qian, and E. R. Dougherty are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: roozbehdn@tamu.edu; xqian@ece.tamu.edu; edward@ece.tamu.edu).

M. Shahrokh Esfahani is with the Division of Oncology and Center for Cancer Systems Biology, Stanford School of Medicine, Stanford University, Stanford, CA 94305 USA (e-mail: shahrokh@stanford.edu).

Digital Object Identifier 10.1109/TSP.2017.2788419

second-order statistics, namely, the covariance matrix, where the posterior distribution is obtained by incorporating observations into the prior distribution of unknown noise parameters. To obtain posterior effective noise statistics, we need a methodology to characterize the relationship between the prior and posterior distributions. In this regard, we use a message passing algorithm based on factor graphs. Factor graphs can be used whenever a global function can factor into a product of local functions [25], [26]. We formulate the likelihood function calculation for unknown noise parameters as a message passing algorithm running in a factor graph and find a closed-form solution for the required update rules. After computing the likelihood function, we generate posterior samples through an MCMC method to compute the posterior effective noise statistics relative to which the OBKF can be designed.

The proposed OBKF method is fairly general and any proper probability distribution can be assumed as the prior distribution for unknown noise parameters. As can be seen later in the simulations section, we perform simulations for two different distributions: uniform distribution and Beta distribution. The final formulation of both the IBR Kalman filter [9] and OBKF are similar, except the need for computing the posterior effective noise statistics in the OBKF. Although we propose a closed-form solution for computing the likelihood function, the computational complexity may still become an issue when the likelihood function is computed for a long sequence of observations being incorporated into the prior distribution or when the state-space model matrices are of extremely large dimensions, which can make matrix calculations such as matrix inversion or matrix determinant time consuming.

The rest of this paper is organized as follows. Section II provides a brief review of the classical and IBR Kalman filters. In Section III, we introduce the proposed OBKF framework. We also lay out the main theorems and show how factor graphs help calculate the posterior effective noise statistics. In Section IV, we apply the OBKF framework to two different applications: target tracking and gene network inference. Finally, we conclude the paper in Section V.

II. Background

Let us first introduce some notations to be employed in the paper. Lowercase and uppercase boldface letters are used to denote vectors and matrices, respectively. The i-th element of vector $\mathbf{v}$ is denoted by $v(i)$. We use $\mathbf{v}_k$ to denote the value of vector $\mathbf{v}$ at time k. For a probability space $(\Omega, \mathcal{E}, P)$, $\mathrm{E}[\bullet]$ and $\mathrm{cov}[\bullet]$ denote the expectation and the covariance matrix of a random vector, respectively. The notations $\mathrm{Tr}(\mathbf{M})$, $\mathbf{M}^T$, and $|\mathbf{M}|$ stand for the trace, transpose, and determinant operators, respectively. A multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

A. Bayesian Robust Filters

Filtering involves a pair of random processes, a *signal process* $\mathbf{x}_k$ and an *observation process* $\mathbf{y}_k$, k being a time index. The aim is to find an estimate $\widehat{\mathbf{x}}_k$ of $\mathbf{x}_k$ through a function (filter) ψ based upon observation process $\mathbf{y}$, i.e., $\widehat{\mathbf{x}}_k = \psi(\mathbf{y}; k)$. The optimal filter $\widehat{\psi}$ is obtained relative to a cost function $C\big(\mathbf{x}_k, \psi(\mathbf{y}; k)\big)$ and a class of filters Ψ:

$$\widehat{\psi}(\mathbf{y}; k) = \arg\min_{\psi \in \Psi} C\big(\mathbf{x}_k, \psi(\mathbf{y}; k)\big). \tag{1}$$

The mean-square error (MSE) is usually used as the cost function, i.e., $C\big(\mathbf{x}_k, \psi(\mathbf{y}; k)\big) = \mathrm{E}\big[\big|\mathbf{x}_k - \psi(\mathbf{y}; k)\big|^2\big]$. In this case, the optimal filter is also called a *minimum mean-square error* (MMSE) filter. Now suppose the statistical model characterizing the relationship between $\mathbf{x}_k$ and $\mathbf{y}_k$ is not fully known and is parameterized by a vector $\boldsymbol{\theta} = [\theta_1, ..., \theta_l]$ of unknown parameters, where $\boldsymbol{\theta} \in \Theta$, Θ being the *uncertainty class*. As discussed in [8], [27], optimal filtering relative to the uncertainty class Θ, cost function C, and filter class Ψ is

$$\psi_{\mathrm{IBR}}(\mathbf{y}; k) = \arg\min_{\psi \in \Psi} \mathrm{E}_{\boldsymbol{\theta}}\big[C_{\boldsymbol{\theta}}\big(\mathbf{x}_k, \psi(\mathbf{y}; k)\big)\big], \tag{2}$$

where the expectation is taken relative to the prior distribution $\pi(\boldsymbol{\theta})$ governing Θ, $C_{\boldsymbol{\theta}}(.)$ characterizes the filter cost relative to $\boldsymbol{\theta}$, and ψ_{IBR} is called an *intrinsically Bayesian robust* (IBR) filter. The term "intrinsically" refers to the fact that the optimality is defined relative to the entire class of filters [8], rather than being constrained to optimal filters within the uncertainty class, as is the case with *model-constrained Bayesian robust* (MCBR) filters [7].

Let $\mathcal{Y}_k = \{\mathbf{y}_0, ..., \mathbf{y}_k\}$ denote the sequence of observations from the underlying model up to time k. In contrary to the optimization for the IBR framework, in this paper we focus on the filter that provides minimum expected cost relative to the posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{Y}_k)$ of the uncertainty class, which is conditioned on observations $\mathcal{Y}_k$. Such a filter is called an *optimal Bayesian filter* (OBF):

$$\psi_{\mathrm{OBF}}(\mathbf{y}; k) = \arg\min_{\psi \in \Psi} \mathrm{E}_{\boldsymbol{\theta}}\big[C_{\boldsymbol{\theta}}\big(\mathbf{x}_k, \psi(\mathbf{y}; k)\big)\big|\mathcal{Y}_k\big]. \tag{3}$$

The concept of designing optimal Bayesian robust operators has been worked out in other problems such as classification where the goal is to design an optimal Bayesian classifier relative to the posterior distribution of the unknown feature-label distribution [28], regression where for a Gaussian model with unknown mean and covariance, optimal Bayesian regression is derived [29], and linear filtering in the context of wide-sense stationary processes where the optimal Bayesian Wiener filter is derived [29]. The terminology "optimal Bayesian Kalman filter" is consistent with these previously treated optimizations relative to the posterior distribution.

B. Kalman Filter

The Kalman filter was first proposed in [12] for discrete time and in [30] for the continuous domain, where it is called the *Kalman-Bucy* filter. The signal-observation model for the Kalman filter is described in terms of a state-space model:

$$\mathbf{x}_{k+1} = \boldsymbol{\Phi}_k \mathbf{x}_k + \boldsymbol{\Gamma}_k \mathbf{u}_k, \tag{4}$$

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k, \tag{5}$$

where (4) and (5) are called the *state equation* and the *observation equation*, respectively. In this model, $\mathbf{x}_k$ and $\mathbf{y}_k$ are of size $n \times 1$ and $m \times 1$ and called the *state vector* and *observation vector*, respectively. $\boldsymbol{\Phi}_k$ is an $n \times n$ matrix called the *state transition matrix*, $\mathbf{u}_k$ is a zero-mean $p \times 1$ random vector called the *process noise vector*, $\boldsymbol{\Gamma}_k$ is of size $n \times p$ and called the *process*

noise transition matrix, $\mathbf{H}_k$ is a matrix of size $m \times n$ called the *observation transition matrix*, and $\mathbf{v}_k$ is a zero-mean random vector of size $m \times 1$ called the *observation noise*. The process and observation noise statistics are summarized below:

$$\begin{aligned} &\mathrm{E}\big[\mathbf{u}_k \mathbf{u}_l^T\big] = \mathbf{Q}\,\delta_{kl},\ \mathrm{E}\big[\mathbf{v}_k \mathbf{v}_l^T\big] = \mathbf{R}\,\delta_{kl} \quad \forall\, k,l = 0,1,... \\ &\mathrm{E}\big[\mathbf{v}_k \mathbf{x}_l^T\big] = \mathbf{0}_{m\times n},\ \mathrm{E}\big[\mathbf{u}_k \mathbf{v}_l^T\big] = \mathbf{0}_{p\times m} \quad \forall\, k,l = 0,1,... \\ &\mathrm{E}\big[\mathbf{u}_k \mathbf{y}_l^T\big] = \mathbf{0}_{p\times m}, \qquad 0 \le l \le k, \end{aligned} \tag{6}$$

where $\mathbf{0}_{p\times m}$ is a zero-valued matrix of size $p \times m$. The Kalman filter provides the estimation $\widehat{\mathbf{x}}_k$ of $\mathbf{x}_k$ based on observations $\mathbf{y}_l$, $l \le k-1$. In the case that $\mathbf{v}_k$ and $\mathbf{u}_k$ are white and Gaussian, the Kalman filter is also the MMSE filter. Otherwise, it is optimal among linear filters.

III. Optimal Bayesian Kalman Filter

Now assume that the covariance matrices of the process and observation noise are not known and parameterized as

$$\mathrm{E}\big[\mathbf{u}_k^{\theta_1} (\mathbf{u}_l^{\theta_1})^T\big] = \mathbf{Q}^{\theta_1}\,\delta_{kl}, \tag{7}$$

$$\mathrm{E}\big[\mathbf{v}_k^{\theta_2} (\mathbf{v}_l^{\theta_2})^T\big] = \mathbf{R}^{\theta_2}\,\delta_{kl}, \tag{8}$$

$\boldsymbol{\theta} = \big[\theta_1, \theta_2\big]$ being the set of unknown parameters governed by the prior distribution $\pi(\boldsymbol{\theta})$. The state-space model belongs to an uncertainty class Θ ($\boldsymbol{\theta} \in \Theta$) of possible state-space models. If θ_1 and θ_2 are statistically independent, then the state-space model can be parameterized as

$$\mathbf{x}_{k+1}^{\theta_1} = \boldsymbol{\Phi}_k \mathbf{x}_k^{\theta_1} + \boldsymbol{\Gamma}_k \mathbf{u}_k^{\theta_1}, \tag{9}$$

$$\mathbf{y}_k^{\boldsymbol{\theta}} = \mathbf{H}_k \mathbf{x}_k^{\theta_1} + \mathbf{v}_k^{\theta_2}. \tag{10}$$

The *intrinsically Bayesian robust* (IBR) Kalman filter that provides optimal performance on average with respect to a prior distribution (consistent with the definition in (2)) has been developed using the notions of *Bayesian orthogonality principle* and *Bayesian innovation process* in [9] and its structure is completely similar to that of the classical Kalman filtering with the noise covariances and the Kalman gain matrix replaced by the expected noise covariances and the effective Kalman gain matrix, respectively.

In distinction to the IBR theory in [9], in this paper optimization is relative to the posterior distribution that incorporates both the prior distribution and observations. Considering the state-space model in (9) and (10), let $\mathcal{Y}_{k-1} = \{\mathbf{y}_0, ..., \mathbf{y}_{k-1}\}$ and $\mathcal{X}_k = \{\mathbf{x}_0, ..., \mathbf{x}_k\}$ be the sequences of observations and states up to times $k-1$ and k, respectively, with $f(\theta, \mathcal{Y}_{k-1}, \mathcal{X}_k)$ being the joint probability distribution of the uncertainty class Θ and observations and states. In the context of optimal Bayesian filtering theory defined in (3), we seek a linear filter of the form

$$\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}} = \sum_{l\le k-1} \mathbf{G}_{k,l}^{\Theta} \mathbf{y}_l^{\boldsymbol{\theta}}, \tag{11}$$

such that

$$\begin{aligned} \mathbf{G}_{k,l}^{\Theta} = \arg \min_{\mathbf{G}_{k,l}\in\mathcal{G}} \mathrm{E}_{\boldsymbol{\theta}} \Bigg[\mathrm{E} \Bigg[\Big(\mathbf{x}_k^{\theta_1} - \sum_{l\le k-1} \mathbf{G}_{k,l} \mathbf{y}_l^{\boldsymbol{\theta}} \Big)^T \\ \times \Big(\mathbf{x}_k^{\theta_1} - \sum_{l\le k-1} \mathbf{G}_{k,l} \mathbf{y}_l^{\boldsymbol{\theta}} \Big) \Bigg] \Big| \mathcal{Y}_{k-1} \Bigg], \end{aligned} \tag{12}$$

where $\mathcal{G}$ is the vector space of all $n \times m$ matrix-valued functions, $\mathbf{G}_{k,l} \in \mathcal{G}$ is a mapping $\mathbf{G}_{k,l} : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{R}^{n\times m}$ such that $\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} ||\mathbf{G}_{k,l}||_2 < \infty$, $||\bullet||_2$ being the L_2 norm and $\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$ computed in (11) is called the *optimal Bayesian least-squares estimate* of $\mathbf{x}_k^{\boldsymbol{\theta}}$.

The following theorem, definition, and lemma are essential for the derivation of the OBKF framework and are restatements of their counterparts in [9] with respect to the posterior distribution. The proofs are similar to those in [9] when written relative to the posterior distribution.

Theorem 1: (*Bayesian orthogonality principle*) The linear estimate $\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$ obtained in (11) is an optimal Bayesian least-squares estimate of $\mathbf{x}_k^{\boldsymbol{\theta}}$, according to (12), if and only if

$$\mathrm{E}_{\boldsymbol{\theta}} \Big[\mathrm{E} \Big[\big(\mathbf{x}_k^{\theta_1} - \widehat{\mathbf{x}}_k^{\boldsymbol{\theta}} \big) \big(\mathbf{y}_l^{\boldsymbol{\theta}} \big)^T \Big] \Big| \mathcal{Y}_{k-1} \Big] = \mathbf{0}_{n\times m} \quad \forall\, l \le k-1. \tag{13}$$

Definition 1: Consider the state-space model in (9) and (10) and let $\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$ be a linear estimate of $\mathbf{x}_k^{\boldsymbol{\theta}}$ that satisfies (11) and (12), then the random process

$$\widetilde{\mathbf{z}}_k^{\boldsymbol{\theta}} = \mathbf{y}_k^{\boldsymbol{\theta}} - \mathbf{H}_k \widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}, \tag{14}$$

is a zero-mean process, called the *Bayesian innovation process*, and $\forall\, l, l' \le k-1$ we have

$$\mathrm{E}_{\boldsymbol{\theta}} \Big[\mathrm{E} \Big[\widetilde{\mathbf{z}}_l^{\boldsymbol{\theta}} \big(\widetilde{\mathbf{z}}_{l'}^{\boldsymbol{\theta}} \big)^T \Big] \big| \mathcal{Y}_{k-1} \Big] = \mathrm{E}_{\boldsymbol{\theta}} \big[\mathbf{H}_l \mathbf{P}_l^{\mathbf{x},\boldsymbol{\theta}} \mathbf{H}_l^T + \mathbf{R}^{\theta_2} \big| \mathcal{Y}_{k-1} \big] \delta_{ll'}, \tag{15}$$

where $\mathbf{P}_l^{\mathbf{x},\boldsymbol{\theta}} = \mathrm{E}[(\mathbf{x}_l^{\theta_1} - \widehat{\mathbf{x}}_l^{\boldsymbol{\theta}})(\mathbf{x}_l^{\theta_1} - \widehat{\mathbf{x}}_l^{\boldsymbol{\theta}})^T]$ is the estimation error covariance matrix of the OBKF at time l relative to $\boldsymbol{\theta}$.

Lemma 1: (*Bayesian information equivalence*) Let $\check{\mathbf{x}}_k^{\boldsymbol{\theta}} = \sum_{l\le k-1} \mathbf{G}_{k,l} \widetilde{\mathbf{z}}_l^{\boldsymbol{\theta}}$ be an estimate of $\mathbf{x}_k^{\boldsymbol{\theta}}$ obtained using the information in $\widetilde{\mathbf{z}}_l^{\boldsymbol{\theta}} = \mathbf{y}_l^{\theta} - \mathbf{H}_k \check{\mathbf{x}}_l^{\boldsymbol{\theta}}$, such that,

$$\mathrm{E}_{\boldsymbol{\theta}} \Big[\mathrm{E} \Big[\big(\mathbf{x}_k^{\theta_1} - \check{\mathbf{x}}_k^{\boldsymbol{\theta}} \big) \big(\widetilde{\mathbf{z}}_l^{\boldsymbol{\theta}} \big)^T \Big] \big| \mathcal{Y}_{k-1} \Big] = \mathbf{0}_{n\times m}. \tag{16}$$

Then $\mathrm{E}_{\boldsymbol{\theta}} \big[\mathrm{E} \big[\big(\mathbf{x}_k^{\theta_1} - \check{\mathbf{x}}_k^{\boldsymbol{\theta}} \big) \big(\mathbf{y}_l^{\boldsymbol{\theta}} \big)^T \big] \big| \mathcal{Y}_{k-1} \big] = \mathbf{0}_{n\times m}$.

Using the Bayesian orthogonality principle and the Bayesian innovation process, the recursive equations constituting the OBKF can be found similarly to those for the IBR Kalman filter in [9]. The block-diagram of the proposed OBKF framework is shown in Fig. 1. We briefly explain how the equations can be derived, referring to [9] for a more complete discussion of the derivations.

According to Lemma 1, we can write $\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$ that satisfies (11) and (12) as:

$$\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}} = \sum_{l\le k-1} \mathbf{G}_{k,l}^{\Theta} \widetilde{\mathbf{z}}_l^{\boldsymbol{\theta}}. \tag{17}$$

As is shown in [9], one can verify that substituting (17) in (16) yields

$$\begin{aligned} \widehat{\mathbf{x}}_k^{\boldsymbol{\theta}} = \sum_{l\le k-1} \mathrm{E}_{\boldsymbol{\theta}} \Big[\mathrm{E} \Big[\mathbf{x}_k^{\theta_1} \big(\widetilde{\mathbf{z}}_l^{\boldsymbol{\theta}} \big)^T \Big] \big| \mathcal{Y}_{k-1} \Big] \mathrm{E}_{\boldsymbol{\theta}}^{-1} \big[\mathbf{H}_l \mathbf{P}_l^{\mathbf{x},\boldsymbol{\theta}} \mathbf{H}_l^T \\ + \mathbf{R}^{\theta_2} \big| \mathcal{Y}_{k-1} \big] \widetilde{\mathbf{z}}_l^{\boldsymbol{\theta}}. \end{aligned} \tag{18}$$

Using (18), an update equation for $\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$ can be found as

$$\widehat{\mathbf{x}}_{k+1}^{\boldsymbol{\theta}} = \boldsymbol{\Phi}_k \widehat{\mathbf{x}}_k^{\boldsymbol{\theta}} + \boldsymbol{\Phi}_k \mathbf{K}_k^{\Theta^*} \widetilde{\mathbf{z}}_k^{\boldsymbol{\theta}}, \tag{19}$$

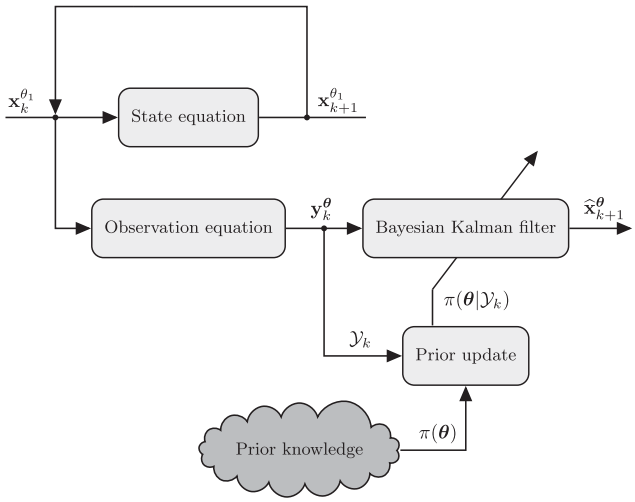

Fig. 1. The schematic representation of the proposed optimal Bayesian Kalman filtering framework.

where

$$\mathbf{K}_k^{\Theta^*} = \mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_{k-1}\big]\mathbf{H}_k^T\,\mathrm{E}_{\boldsymbol{\theta}}^{-1}\big[\mathbf{H}_k\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\mathbf{H}_k^T + \mathbf{R}^{\theta_2}\big|\mathcal{Y}_{k-1}\big], \tag{20}$$

is the *posterior effective Kalman gain* matrix. Note that we use $\mathbf{K}_k^{\Theta^*}$ and $\mathbf{K}_k^{\Theta}$ to distinguish between the effective Kalman gain matrix obtained relative to the posterior distribution in this paper and the one obtained relative to the prior distribution in [9].

Letting $\mathbf{x}_k^{e,\boldsymbol{\theta}} = \mathbf{x}_k^{\theta_1} - \widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$ be the Bayesian least-squares estimation error at time k, the update equation for $\mathbf{x}_k^{e,\boldsymbol{\theta}}$ is

$$\mathbf{x}_{k+1}^{e,\boldsymbol{\theta}} = \boldsymbol{\Phi}_k\big(\mathbf{I} - \mathbf{K}_k^{\Theta^*}\mathbf{H}_k\big)\mathbf{x}_k^{e,\boldsymbol{\theta}} + \boldsymbol{\Gamma}_k\mathbf{u}_k^{\theta_1} - \boldsymbol{\Phi}_k\mathbf{K}_k^{\Theta^*}\mathbf{v}_k^{\theta_2}. \tag{21}$$

Letting $\mathbf{P}_{k+1}^{\mathbf{x},\boldsymbol{\theta}} = \mathbf{E}\big[\mathbf{x}_{k+1}^{e,\boldsymbol{\theta}}(\mathbf{x}_{k+1}^{e,\boldsymbol{\theta}})^T\big]$ and after some mathematical manipulations,

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_{k+1}^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big] &= \boldsymbol{\Phi}_k\big(\mathbf{I} - \mathbf{K}_k^{\Theta^*}\mathbf{H}_k\big)\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big]\boldsymbol{\Phi}_k^T \\ &\quad + \boldsymbol{\Gamma}_k\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]\boldsymbol{\Gamma}_k^T. \end{aligned}\tag{22}$$

This completes the required recursive equations needed for the proposed OBKF framework. Note that in the right hand side of (22) we have $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big]$; however, what we have from the last recursion is in fact $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_{k-1}\big]$. There are two options to handle this issue. The first option is to do the recursive updates in (20) and (22) at each time k from the beginning using $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{R}^{\theta_2}\big|\mathcal{Y}_k\big]$, i.e., first compute $\mathbf{K}_0^{\Theta^*}$ using $\mathrm{cov}[\mathbf{x}_0]$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{R}^{\theta_2}\big|\mathcal{Y}_k\big]$, then compute $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_1^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big]$ using $\mathbf{K}_0^{\Theta^*}$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]$, then compute $\mathbf{K}_1^{\Theta^*}$ using $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_1^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big]$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{R}^{\theta_2}\big|\mathcal{Y}_k\big]$, and so on, until we reach $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big]$. The other option is to make the approximation $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big] \approx \mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_{k-1}\big]$ and use $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_{k-1}\big]$ obtained from the last recursion in place of $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_k\big]$ in (22). This option is computationally more efficient as we do not need to repeat all the recursions at each time k from the beginning. We used the second option in our simulations.

Table I compares the recursive equations for the classical Kalman filter and the proposed OBKF framework. Both filters have similar structures with the original noise covariances $\mathbf{R}$ and $\mathbf{Q}$ for the classical Kalman filter replaced by the posterior effective noise statistics $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{R}^{\theta_2}\big|\mathcal{Y}_k\big]$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]$ for the OBKF framework, respectively. In other words, an optimal Bayesian Kalman filter is a classical Kalman filter designed relative to the posterior effective noise statistics.

A. Calculation of Posterior Effective Noise Statistics

To implement an OBKF, we need to compute the conditional expectations $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{R}^{\theta_2}\big|\mathcal{Y}_k\big]$ with respect to the posterior distribution $\pi(\boldsymbol{\theta}\big|\mathcal{Y}_k) \propto f(\mathcal{Y}_k\big|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where $f(\mathcal{Y}_k\big|\boldsymbol{\theta})$ is the likelihood function of $\boldsymbol{\theta}$ given the sequence of observations $\mathcal{Y}_k$. As there is no closed-form solution for $\pi(\boldsymbol{\theta}\big|\mathcal{Y}_k)$ for many prior distributions, we employ an MCMC method to generate samples from the posterior distribution $\pi(\boldsymbol{\theta}\big|\mathcal{Y}_k)$ and then approximate $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{R}^{\theta_2}\big|\mathcal{Y}_k\big]$ as sample means of the generated MCMC samples. First we need to compute the likelihood function $f(\mathcal{Y}_k\big|\boldsymbol{\theta})$.

The Markov assumption in the state-space model postulates that, given the current state $\mathbf{x}_k$, the next state $\mathbf{x}_{k+1}$ and the current observation $\mathbf{y}_k$ are normally distributed and possess the following conditional independence properties:

$$f\big(\mathbf{y}_k\big|\mathcal{Y}_{k-1},\mathcal{X}_k,\boldsymbol{\theta}\big) = f\big(\mathbf{y}_k\big|\mathbf{x}_k,\boldsymbol{\theta}\big) = \mathcal{N}(\mathbf{y}_k;\mathbf{H}_k\mathbf{x}_k,\mathbf{R}^{\theta_2}), \tag{23}$$

$$\begin{aligned} f\big(\mathbf{x}_{k+1}\big|\mathcal{Y}_{k+1},\mathcal{X}_k,\boldsymbol{\theta}\big) &= f\big(\mathbf{x}_{k+1}\big|\mathbf{x}_k,\boldsymbol{\theta}\big) \\ &= \mathcal{N}\big(\mathbf{x}_{k+1};\boldsymbol{\Phi}_k\mathbf{x}_k,\widetilde{\mathbf{Q}}_k^{\theta_1}\big), \end{aligned}\tag{24}$$

where $\widetilde{\mathbf{Q}}_k^{\theta_1} = \boldsymbol{\Gamma}_k\mathbf{Q}^{\theta_1}\boldsymbol{\Gamma}_k^T$. To compute $f(\mathcal{Y}_k\big|\boldsymbol{\theta})$, we first write it as the marginalization of $f(\mathcal{Y}_k,\mathcal{X}_k\big|\boldsymbol{\theta})$ over $\mathcal{X}_k$ and then factorize $f(\mathcal{Y}_k,\mathcal{X}_k\big|\boldsymbol{\theta})$ by taking advantage of the Markov assumptions given in (23) and (24):

$$\begin{aligned} &f(\mathcal{Y}_k\big|\boldsymbol{\theta}) \\ &= \underbrace{\int\cdots\int}_{\mathbf{x}_0,\ldots,\mathbf{x}_k} f(\mathcal{Y}_k,\mathcal{X}_k\big|\boldsymbol{\theta})\,d\mathbf{x}_0\ldots d\mathbf{x}_k \\ &= \underbrace{\int\cdots\int}_{\mathbf{x}_0,\ldots,\mathbf{x}_k} f(\mathbf{y}_k|\mathbf{x}_k,\boldsymbol{\theta})f(\mathcal{Y}_{k-1},\mathcal{X}_k\big|\boldsymbol{\theta})\,d\mathbf{x}_0\ldots d\mathbf{x}_k \\ &= \underbrace{\int\cdots\int}_{\mathbf{x}_0,\ldots,\mathbf{x}_k} \Big\{f(\mathbf{y}_k|\mathbf{x}_k,\boldsymbol{\theta})f(\mathbf{x}_k\big|\mathbf{x}_{k-1},\boldsymbol{\theta}) \\ &\qquad \times\ f(\mathcal{Y}_{k-1},\mathcal{X}_{k-1}|\boldsymbol{\theta})\Big\}d\mathbf{x}_0\ldots d\mathbf{x}_k \\ &\ \ \vdots \\ &= \underbrace{\int\cdots\int}_{\mathbf{x}_0,\ldots,\mathbf{x}_k} \prod_{i=0}^{k} f(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})\prod_{i=1}^{k} f(\mathbf{x}_i\big|\mathbf{x}_{i-1},\boldsymbol{\theta})f(\mathbf{x}_0)\,d\mathbf{x}_0\ldots d\mathbf{x}_k, \end{aligned}\tag{25}$$

where in the second and third equalities, being obtained from the chain rule, we used (23) and (24), respectively.

TABLE I
COMPARISON OF THE RECURSIVE EQUATIONS REQUIRED FOR THE CLASSICAL AND OPTIMAL BAYESIAN KALMAN FILTERS

Classical Kalman Filter	Optimal Bayesian Kalman Filter (OBKF)
$\widetilde{\mathbf{z}}_k = \mathbf{y}_k - \mathbf{H}_k\widehat{\mathbf{x}}_k$	$\widetilde{\mathbf{z}}_k^{\boldsymbol{\theta}} = \mathbf{y}_k^{\boldsymbol{\theta}} - \mathbf{H}_k\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$
$\mathbf{K}_k = \mathbf{P}_k^{\mathbf{x}}\mathbf{H}_k^T\left(\mathbf{H}_k\mathbf{P}_k^{\mathbf{x}}\mathbf{H}_k^T + \mathbf{R}\right)^{-1}$	$\mathbf{K}_k^{\Theta^*} = \mathrm{E}_{\boldsymbol{\theta}}\left[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\middle\vert\mathcal{Y}_{k-1}\right]\mathbf{H}_k^T\mathrm{E}_{\boldsymbol{\theta}}^{-1}\left[\mathbf{H}_k\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\mathbf{H}_k^T + \mathbf{R}^{\theta_2}\middle\vert\mathcal{Y}_{k-1}\right]$
$\widehat{\mathbf{x}}_{k+1} = \boldsymbol{\Phi}_k\widehat{\mathbf{x}}_k + \boldsymbol{\Phi}_k\mathbf{K}_k\widetilde{\mathbf{z}}_k$	$\widehat{\mathbf{x}}_{k+1}^{\boldsymbol{\theta}} = \boldsymbol{\Phi}_k\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}} + \boldsymbol{\Phi}_k\mathbf{K}_k^{\Theta^*}\widetilde{\mathbf{z}}_k^{\boldsymbol{\theta}}$
$\mathbf{P}_{k+1}^{\mathbf{x}} = \boldsymbol{\Phi}_k\left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k\right)\mathbf{P}_k^{\mathbf{x}}\boldsymbol{\Phi}_k^T + \boldsymbol{\Gamma}_k\mathbf{Q}\boldsymbol{\Gamma}_k^T$	$\mathrm{E}_{\boldsymbol{\theta}}\left[\mathbf{P}_{k+1}^{\mathbf{x},\boldsymbol{\theta}}\middle\vert\mathcal{Y}_k\right] = \boldsymbol{\Phi}_k\left(\mathbf{I} - \mathbf{K}_k^{\Theta^*}\mathbf{H}_k\right)\mathrm{E}_{\boldsymbol{\theta}}\left[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\middle\vert\mathcal{Y}_k\right]\boldsymbol{\Phi}_k^T + \boldsymbol{\Gamma}_k\mathrm{E}_{\boldsymbol{\theta}}\left[\mathbf{Q}^{\theta_1}\middle\vert\mathcal{Y}_k\right]\boldsymbol{\Gamma}_k^T$

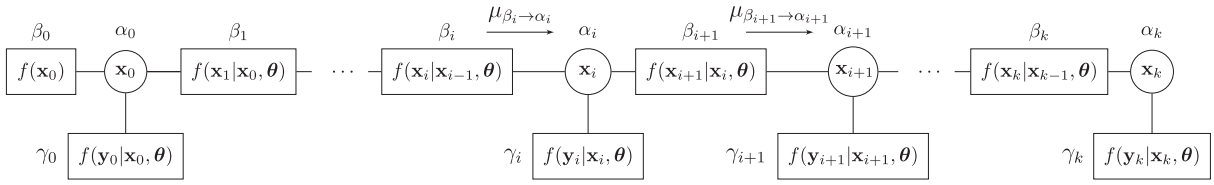

Fig. 2. The factor graph model utilized to compute the likelihood function $f(\mathcal{Y}_k|\boldsymbol{\theta})$.

The relation in (25) can be regarded as the factorization of a global function for which we can employ the sum-product algorithm, which is a message-passing algorithm. To visualize the computations in a sum-product algorithm, a bipartite graph, called the *factor graph*, can be associated [25].

A factor graph shows how a global function can factor as the product of local functions. Factor graphs have a specific configuration. Let $f(X) = \prod_{i\in I} f_i(X_i)$, I being a finite set of discrete indices, be the factorization of a global function $f(X)$ with variables $X = \{x_1, ..., x_N\}$ into a product of local functions $f_i(X_i)$, X_i being a subset of X. The factor graph possesses a *variable node* x_j and a *factor node* f_i corresponding to each variable x_j of the global function and each local function $f_i(X_i)$, respectively. If $x_j \in X_i$, an edge exists between the variable node x_j and the factor node f_i. Let $\mu_{v_i\to v_j}$ denote the message corresponding to a specific function sent from node v_i to v_j. To use a factor graph for a sum-product algorithm, each variable node x sends the product of all messages it receives from its children nodes as a new message to its parent nodes, i.e., if p_x be a parent node for x, $\mu_{x\to p_x} = \prod_{v\in ch(x)}\mu_{v\to x}$, $ch(x)$ being the set of all children nodes for x. For a non-leaf factor node f_i, it computes the product of all incoming messages from its children nodes with its local function and then marginalizes out all variables other than x to send a message to its parent variable node x. In other words,

$$\mu_{f_i\to x} = \int_{(X_i\backslash x)} f_i(X_i)\prod_{v\in ch(f_i)}\mu_{v\to f_i}\,d(X_i\backslash x).$$

A node in the factor graph operates when it receives all messages from its children nodes. The first step for running a factor graph is that each leaf function node sends the message corresponding to its local function to its parent nodes.

Fig. 2 shows the factor graph model adopted to compute the likelihood function in (25). In this factor graph, the variable node corresponding to $\mathbf{x}_i$, the factor node corresponding to $f(\mathbf{x}_i|\mathbf{x}_{i-1},\boldsymbol{\theta})$, and the factor node corresponding to $f(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})$ are named as nodes α_i, β_i, and γ_i, respectively. Messages transmitted in the factor graph contain three components: the scale, mean vector, and the covariance matrix of a scaled multivariate Gaussian function. For example, the message $\mu_{\beta_i\to\alpha_i} = (S_i, \mathbf{M}_i, \boldsymbol{\Sigma}_i)$ sent from β_i to α_i corresponds to the scaled Gaussian function $S_i\,\mathcal{N}(\mathbf{x}_i;\mathbf{M}_i,\boldsymbol{\Sigma}_i)$. Also the message sent from γ_i to α_i is $\mu_{\gamma_i\to\alpha_i} = (1, \mathbf{H}_i\mathbf{x}_i, \mathbf{R}^{\theta_2})$, which corresponds to $\mathcal{N}(\mathbf{y}_i;\mathbf{H}_i\mathbf{x}_i,\mathbf{R}^{\theta_2})$. Note that since we want to compare the likelihood functions of different MCMC samples generated for $\boldsymbol{\theta}$ and, as shown later in the lemma, the scale parameter depends on $\boldsymbol{\theta}$, we need to compute the scale parameter for the likelihood function.

Assume that node α_i has received message $\mu_{\beta_i\to\alpha_i} = (S_i, \mathbf{M}_i, \boldsymbol{\Sigma}_i)$. Now we aim to compute the outgoing message $\mu_{\beta_{i+1}\to\alpha_{i+1}}$ from node β_{i+1} to node α_{i+1}. Computing $\mu_{\beta_{i+1}\to\alpha_{i+1}}$ corresponds to the computation of the following integral:

$$\int_{\mathbf{x}_i}\mathcal{N}\Big(\mathbf{x}_{i+1};\boldsymbol{\Phi}_i\mathbf{x}_i,\widetilde{\mathbf{Q}}_i^{\theta_1}\Big)\mathcal{N}(\mathbf{y}_i;\mathbf{H}_i\mathbf{x}_i,\mathbf{R}^{\theta_2}) \times S_i\,\mathcal{N}(\mathbf{x}_i;\mathbf{M}_i,\boldsymbol{\Sigma}_i)d\mathbf{x}_i. \tag{26}$$

This integral can be computed using the following lemma.

Lemma 2: The solution of the integral given in (26) is a scaled multivariate Gaussian function $S_{i+1}\,\mathcal{N}(\mathbf{x}_{i+1};\mathbf{M}_{i+1}, \boldsymbol{\Sigma}_{i+1})$, whose parameters S_{i+1}, $\mathbf{M}_{i+1}$, and $\boldsymbol{\Sigma}_{i+1}$ are given by

$$\boldsymbol{\Sigma}_{i+1}^{-1} = (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1} - (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\boldsymbol{\Phi}_i\boldsymbol{\Lambda}_i\boldsymbol{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}, \tag{27}$$

$$\mathbf{M}_{i+1} = \boldsymbol{\Sigma}_{i+1}(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\boldsymbol{\Phi}_i\boldsymbol{\Lambda}_i\Big(\mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i + \boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i\Big), \tag{28}$$

$$S_{i+1} = S_i\sqrt{\frac{\left|\boldsymbol{\Lambda}_i\right|\left|\boldsymbol{\Sigma}_{i+1}\right|}{\left|\widetilde{\mathbf{Q}}_i^{\theta_1}\right|\left|\boldsymbol{\Sigma}_i\right|}}\mathcal{N}\big(\mathbf{y}_i;\mathbf{0}_{m\times 1},\mathbf{R}^{\theta_2}\big) \times \exp\left(\frac{\mathbf{M}_{i+1}^T\boldsymbol{\Sigma}_{i+1}^{-1}\mathbf{M}_{i+1} + \mathbf{W}_i^T\boldsymbol{\Lambda}_i\mathbf{W}_i - \mathbf{M}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i}{2}\right), \tag{29}$$

where $\mathbf{W}_i = \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i + \mathbf{\Sigma}_i^{-1}\mathbf{M}_i$ and

$$\mathbf{\Lambda}_i = \left(\mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{\Phi}_i + \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_i + \mathbf{\Sigma}_i^{-1}\right)^{-1}. \quad (30)$$

Proof: Refer to Appendix. ■

The update rules given in Lemma 2 should be iterated for $0 \le i \le k-1$ to finally obtain the message $\mu_{\beta_k \to \alpha_k} = (S_k, \mathbf{M}_k, \mathbf{\Sigma}_k)$. Then the likelihood function is obtained as:

$$\begin{aligned} f(\mathcal{Y}_k|\boldsymbol{\theta}) &= \int_{\mathbf{x}_k} \mathcal{N}(\mathbf{y}_k; \mathbf{H}_k\mathbf{x}_k, \mathbf{R}^{\theta_2}) S_k\, \mathcal{N}(\mathbf{x}_k; \mathbf{M}_k, \mathbf{\Sigma}_k) d\mathbf{x}_k \\ &= \int_{\mathbf{x}_k} \frac{S_k}{\sqrt{(2\pi)^m|\mathbf{R}^{\theta_2}|}\sqrt{(2\pi)^n|\mathbf{\Sigma}_k|}} \\ &\quad \times \exp\left(\frac{-1}{2}\Big(\big(\mathbf{y}_k - \mathbf{H}_k\mathbf{x}_k\big)^T(\mathbf{R}^{\theta_2})^{-1}\big(\mathbf{y}_k - \mathbf{H}_k\mathbf{x}_k\big)\right. \\ &\quad \left. + \big(\mathbf{x}_k - \mathbf{M}_k\big)^T\mathbf{\Sigma}_k^{-1}\big(\mathbf{x}_k - \mathbf{M}_k\big)\Big)\right) d\mathbf{x}_k. \end{aligned} \quad (31)$$

Similar to the procedure in Appendix, in the exponent of (31) we complete the square for $\mathbf{x}_k$. Letting

$$\mathbf{\Delta}_k^{-1} = \mathbf{H}_k^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_k + \mathbf{\Sigma}_k^{-1}, \quad (32)$$

$$\mathbf{G}_k = \mathbf{\Delta}_k\Big(\mathbf{H}_k^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_k + \mathbf{\Sigma}_k^{-1}\mathbf{M}_k\Big), \quad (33)$$

we rewrite (31) as

$$\begin{aligned} f(\mathcal{Y}_k|\boldsymbol{\theta}) &= \int_{\mathbf{x}_k} \frac{S_k}{\sqrt{(2\pi)^m|\mathbf{R}^{\theta_2}|}\sqrt{(2\pi)^n|\mathbf{\Sigma}_k|}} \\ &\quad \times \exp\left(\frac{-1}{2}\Big(\big(\mathbf{x}_k - \mathbf{G}_k\big)^T\mathbf{\Delta}_k^{-1}\big(\mathbf{x}_k - \mathbf{G}_k\big)\right. \\ &\quad \left. - \mathbf{G}_k^T\mathbf{\Delta}_k^{-1}\mathbf{G}_k + \mathbf{y}_k^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_k + \mathbf{M}_k^T\mathbf{\Sigma}_k^{-1}\mathbf{M}_k\Big)\right) d\mathbf{x}_k \\ &= \int_{\mathbf{x}_k} S_k\sqrt{\frac{|\mathbf{\Delta}_k|}{|\mathbf{\Sigma}_k|}}\mathcal{N}(\mathbf{x}_k; \mathbf{G}_k, \mathbf{\Delta}_k)\mathcal{N}\big(\mathbf{y}_k; \mathbf{0}_{m\times 1}, \mathbf{R}^{\theta_2}\big) \\ &\quad \times \exp\Big(\frac{1}{2}\big(\mathbf{G}_k^T\mathbf{\Delta}_k^{-1}\mathbf{G}_k - \mathbf{M}_k^T\mathbf{\Sigma}_k^{-1}\mathbf{M}_k\big)\Big) d\mathbf{x}_k \\ &= S_k\sqrt{\frac{|\mathbf{\Delta}_k|}{|\mathbf{\Sigma}_k|}}\mathcal{N}\big(\mathbf{y}_k; \mathbf{0}_{m\times 1}, \mathbf{R}^{\theta_2}\big) \\ &\quad \times \exp\Big(\frac{1}{2}\big(\mathbf{G}_k^T\mathbf{\Delta}_k^{-1}\mathbf{G}_k - \mathbf{M}_k^T\mathbf{\Sigma}_k^{-1}\mathbf{M}_k\big)\Big). \end{aligned} \quad (34)$$

Hence, using the adopted sum-product and factor graph algorithm, the likelihood function $f(\mathcal{Y}_k|\boldsymbol{\theta})$ can be obtained according to (34), where S_k, $\mathbf{M}_k$, and $\mathbf{\Sigma}_k$ are computed recursively using (29), (28), and (27), respectively. $\mathbf{\Delta}_k$ and $\mathbf{G}_k$ are obtained according to (32) and (33), respectively.

To estimate the posterior effective noise statistics $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]$ and $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{R}^{\theta_2}\big|\mathcal{Y}_k\big]$, we employ the Metropolis Hastings MCMC [31]. Let the last accepted MCMC sample in the sequence of samples be $\boldsymbol{\theta}^{(j)}$ generated at the j-th iteration. A candidate MCMC sample $\boldsymbol{\theta}^{\text{candid}}$ will be drawn according to a proposal distribution $f(\boldsymbol{\theta}^{\text{candid}}|\boldsymbol{\theta}^{(j)})$. The candidate MCMC sample $\boldsymbol{\theta}^{\text{candid}}$ will be either accepted or rejected according to an acceptance ratio r defined as

$$\begin{aligned} r &= \min\left\{1, \frac{f\big(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{\text{candid}}\big)f\big(\mathcal{Y}_k|\boldsymbol{\theta}^{\text{candid}}\big)\pi(\boldsymbol{\theta}^{\text{candid}})}{f\big(\boldsymbol{\theta}^{\text{candid}}|\boldsymbol{\theta}^{(j)}\big)f\big(\mathcal{Y}_k|\boldsymbol{\theta}^{(j)}\big)\pi(\boldsymbol{\theta}^{(j)})}\right\} \\ &= \min\left\{1, \frac{f\big(\mathcal{Y}_k|\boldsymbol{\theta}^{\text{candid}}\big)\pi(\boldsymbol{\theta}^{\text{candid}})}{f\big(\mathcal{Y}_k|\boldsymbol{\theta}^{(j)}\big)\pi(\boldsymbol{\theta}^{(j)})}\right\}, \end{aligned} \quad (35)$$

where the second formula is used when the proposal distribution is symmetric, i.e., $f(\boldsymbol{\theta}^{\text{candid}}|\boldsymbol{\theta}^{(j)}) = f(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{\text{candid}})$. The $(j+1)$-th MCMC sample is

$$\boldsymbol{\theta}^{(j+1)} = \begin{cases} \boldsymbol{\theta}^{\text{candid}} & \text{with probability } r \\ \boldsymbol{\theta}^{(j)} & \text{otherwise} \end{cases}.$$

Repeating this process of drawing samples based on the proposal distribution and accepting or rejecting them based on the acceptance ratio generates a sequence of MCMC samples. The positivity of the proposal distribution ($f(\boldsymbol{\theta}^{\text{candid}}|\boldsymbol{\theta}^{(j)}) > 0$ for any $\boldsymbol{\theta}^{(j)}$) is a sufficient condition for having an ergodic Markov chain of MCMC samples whose steady-state distribution is the target distribution $\pi(\boldsymbol{\theta}|\mathcal{Y}_k)$ [32]. After generating enough MCMC samples, the posterior effective noise statistics can be approximated by computing the sample mean of the accepted MCMC samples. For the simulations in this paper, we use a Gaussian distribution as the proposal distribution in the MCMC step.

The steps for implementation of the proposed optimal Bayesian Kalman filtering framework are summarized in Algorithms 1, 2, and 3. Algorithm 1 shows the filtering procedure based on the proposed OBKF framework. As mentioned in line 10 of Algorithm 1, at each iteration we should compute the posterior effective noise statistics using the MCMC method. The pseudo-code for the MCMC step is given in Algorithm 2. The likelihood function of each generated MCMC sample is computed using the factor-graph-based step outlined in Algorithm 3.

Algorithm 1: Optimal Bayesian Kalman Filter.

1: **input**: $\pi(\boldsymbol{\theta})$, $\mathbf{\Phi}_k$, $\mathbf{H}_k$, $\mathbf{Q}^{\theta_1}$, $\mathbf{R}^{\theta_2}$, $\mathbf{\Gamma}_k$, $\mathcal{Y}_k$
2: **output**: $\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$
3: $\widehat{\mathbf{x}}_0^{\boldsymbol{\theta}} \leftarrow \mathrm{E}[\mathbf{x}_0]$
4: $\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_0^{\mathbf{x},\boldsymbol{\theta}}|\mathcal{Y}_{-1}\big] \leftarrow \mathrm{cov}[\mathbf{x}_0]$
5: $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{Q}^{\theta_1}|\mathcal{Y}_{-1}] \leftarrow \mathrm{E}_{\theta_1}[\mathbf{Q}^{\theta_1}]$, $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{R}^{\theta_2}|\mathcal{Y}_{-1}] \leftarrow \mathrm{E}_{\theta_2}[\mathbf{R}^{\theta_2}]$
6: $k \leftarrow 0$
7: **for** $k = 1, 2, \ldots$ **do**
8: $\quad \widetilde{\mathbf{z}}_k^{\boldsymbol{\theta}} \leftarrow \mathbf{y}_k^{\boldsymbol{\theta}} - \mathbf{H}_k\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}}$
9: $\quad \mathbf{K}_k^{\Theta^*} \leftarrow \mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\big|\mathcal{Y}_{k-1}\big]\mathbf{H}_k^T\mathrm{E}_{\boldsymbol{\theta}}^{-1}\big[\mathbf{H}_k\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}\mathbf{H}_k^T + \mathbf{R}^{\theta_2}\big|\mathcal{Y}_{k-1}\big]$
10: $\quad \mathrm{E}_{\boldsymbol{\theta}}[\mathbf{Q}^{\theta_1}|\mathcal{Y}_k]$, $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{R}^{\theta_2}|\mathcal{Y}_k] \leftarrow \mathrm{MCMC}(\mathcal{Y}_k, \pi(\boldsymbol{\theta}))$
11: $\quad \mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_{k+1}^{\mathbf{x},\boldsymbol{\theta}}|\mathcal{Y}_k\big] \leftarrow \mathbf{\Phi}_k\big(\mathbf{I} - \mathbf{K}_k^{\Theta^*}\mathbf{H}_k\big)\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{P}_k^{\mathbf{x},\boldsymbol{\theta}}|\mathcal{Y}_{k-1}\big]\mathbf{\Phi}_k^T + \mathbf{\Gamma}_k\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]\mathbf{\Gamma}_k^T + \mathbf{\Gamma}_k\mathrm{E}_{\boldsymbol{\theta}}\big[\mathbf{Q}^{\theta_1}\big|\mathcal{Y}_k\big]\mathbf{\Gamma}_k^T$
12: $\quad \widehat{\mathbf{x}}_{k+1}^{\boldsymbol{\theta}} \leftarrow \mathbf{\Phi}_k\widehat{\mathbf{x}}_k^{\boldsymbol{\theta}} + \mathbf{\Phi}_k\mathbf{K}_k^{\Theta^*}\widetilde{\mathbf{z}}_k^{\boldsymbol{\theta}}$
$\quad$ **return** $\widehat{\mathbf{x}}_{k+1}^{\boldsymbol{\theta}}$
13: **end for**

Algorithm 2: MCMC Computation.

function MCMC $(\mathcal{Y}_k, \pi(\boldsymbol{\theta}))$
  $\boldsymbol{\theta}^{(0)} \sim \pi(\boldsymbol{\theta})$
  $i \leftarrow 1$
  while $i \leq$ num_iterations **do**
    $\boldsymbol{\theta}^{\text{candid}} \sim f(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)})$
    $f(\boldsymbol{\theta}^{\text{candid}}|\mathcal{Y}_k) \leftarrow$ FACTOR GRAPH $(\boldsymbol{\theta}, \mathcal{Y}_k)$
    $r \leftarrow \min\left\{1, \frac{f\left(\mathcal{Y}_k|\boldsymbol{\theta}^{\text{candid}}\right)\pi(\boldsymbol{\theta}^{\text{candid}})}{f\left(\mathcal{Y}_k|\boldsymbol{\theta}^{(i-1)}\right)\pi(\boldsymbol{\theta}^{(i-1)})}\right\}$
    $\zeta \sim \text{unif}(0,1)$
    if $\zeta < r$ **then**
      $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{\text{candid}}$
    else
      $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{(i-1)}$
    $i \leftarrow i+1$
  return $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}\}$

Algorithm 3: Factor-Graph-Based Likelihood Function Calculation.

function FACTOR GRAPH $(\boldsymbol{\theta}, \mathcal{Y}_k)$
  $\mathbf{M}_0 \leftarrow \mathrm{E}[\mathbf{x}_0]$
  $S_0 \leftarrow 1$
  $\boldsymbol{\Sigma}_0 \leftarrow \text{cov}[\mathbf{x}_0]$
  $i \leftarrow 0$
  while $i \leq k-1$ **do**
    $\mathbf{W}_i \leftarrow \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i + \boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i$
    $\boldsymbol{\Lambda}_i^{-1} \leftarrow \boldsymbol{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\boldsymbol{\Phi}_i + \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_i + \boldsymbol{\Sigma}_i^{-1}$
    $\boldsymbol{\Sigma}_{i+1}^{-1} \leftarrow (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1} - (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\boldsymbol{\Phi}_i\boldsymbol{\Lambda}_i\boldsymbol{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}$
    $\mathbf{M}_{i+1} \leftarrow \boldsymbol{\Sigma}_{i+1}(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\boldsymbol{\Phi}_i\boldsymbol{\Lambda}_i\left(\mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i + \boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i\right)$
    $S_{i+1} \leftarrow$ using (29)
    $i \leftarrow i+1$
  $\boldsymbol{\Delta}_k^{-1} \leftarrow \mathbf{H}_k^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_k + \boldsymbol{\Sigma}_k^{-1}$
  $\mathbf{G}_k \leftarrow \boldsymbol{\Delta}_k\left(\mathbf{H}_k^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_k + \boldsymbol{\Sigma}_k^{-1}\mathbf{M}_k\right)$
  $f(\mathcal{Y}_k|\boldsymbol{\theta}) \leftarrow$ using (34)
    return $f(\mathcal{Y}_k|\boldsymbol{\theta})$

IV. Simulation Results and Performance Analysis

In this section, we demonstrate the performance of the proposed optimal Bayesian Kalman filtering framework by considering two different state-space models used for target tracking and gene regulatory network inference problems. We compare the performance of the proposed OBKF with four other Kalman filtering approaches: optimal model-specific, minimax, IBR Kalman, and the maximum *a posteriori* (MAP) approach. The optimal model-specific Kalman filter is the classical Kalman filter designed relative to the underlying true noise parameters. The minimax Kalman filter has the best worst-case performance across the uncertainty class and is obtained as

$$\boldsymbol{\theta}_{\text{mm}} = \arg\min_{\boldsymbol{\theta}' \in \Theta} \max_{\boldsymbol{\theta} \in \Theta} \lim_{k \to \infty} \text{Tr}\left(\mathbf{P}_k^{\mathbf{x}}(\boldsymbol{\theta}; \boldsymbol{\theta}')\right),$$

where $\mathbf{P}_k^{\mathbf{x}}(\boldsymbol{\theta}; \boldsymbol{\theta}') = \mathrm{E}\left[\mathbf{x}_k^e(\boldsymbol{\theta}; \boldsymbol{\theta}')(\mathbf{x}_k^e(\boldsymbol{\theta}; \boldsymbol{\theta}'))^T\right]$ is the estimation error covariance matrix resulting from applying a Kalman filter designed relative to model $\boldsymbol{\theta}' = [\theta_1', \theta_2']$ to model $\boldsymbol{\theta} = [\theta_1, \theta_2]$ and can be calculated as [9]

$$\mathbf{P}_{k+1}^{\mathbf{x}}(\boldsymbol{\theta}; \boldsymbol{\theta}') = \boldsymbol{\Phi}_k\left(\mathbf{I} - \mathbf{K}_k^{\boldsymbol{\theta}'}\mathbf{H}_k\right)\mathbf{P}_k^{\mathbf{x}}(\boldsymbol{\theta}; \boldsymbol{\theta}')\left(\mathbf{I} - \mathbf{K}_k^{\boldsymbol{\theta}'}\mathbf{H}_k\right)^T\boldsymbol{\Phi}_k^T + \boldsymbol{\Gamma}_k\mathbf{Q}^{\theta_1}\boldsymbol{\Gamma}_k^T + \boldsymbol{\Phi}_k\mathbf{K}_k^{\boldsymbol{\theta}'}\mathbf{R}^{\theta_2}\left(\mathbf{K}_k^{\boldsymbol{\theta}'}\right)^T\boldsymbol{\Phi}_k^T, \quad (36)$$

where $\mathbf{K}_k^{\boldsymbol{\theta}'}$ is the Kalman gain matrix of the classical Kalman filter relative to $\boldsymbol{\theta}'$. The IBR Kalman filter in [9] is the Kalman filter with the optimal performance on average relative to the prior distribution $\pi(\boldsymbol{\theta})$. Finally, for the MAP approach, we design a model-specific Kalman filter relative to the MAP estimates of the noise parameters. To find the MAP estimates, we generate a number of samples from the prior distribution and then calculate the posterior probability for each sample via the likelihood function computed using the proposed factor-graph-based method. The sample with the largest *a posteriori* probability is chosen as the MAP estimate and used for designing the Kalman filter.

A. Example: Target Tracking

Consider the tracking problem in 2-dimensional space. The state vector is given by $\mathbf{x}_k = [p_x\ v_x\ p_y\ v_y]^T$, where p_x, v_x, and p_y, v_y are the position and velocity in the x and y dimensions, respectively. Assume that the speed is constant except for small random perturbations due to maneuvers and measurements are obtained every τ seconds. The state-space model describing the object dynamic equation is of form (4) and (5) with the following matrices [33], [34]:

$$\boldsymbol{\Phi}_k = \begin{bmatrix} 1 & \tau & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H}_k = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{\Gamma}_k = \mathbf{I}.$$

The process noise and observation noise covariance matrices are

$$\mathbf{Q} = q \times \begin{bmatrix} \tau^3/3 & \tau^2/2 & 0 & 0 \\ \tau^2/2 & \tau & 0 & 0 \\ 0 & 0 & \tau^3/3 & \tau^2/2 \\ 0 & 0 & \tau^2/2 & \tau \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix},$$

where q is the process noise intensity that governs the velocity deviation in each dimension. We let $\tau = 1$ second. The initial conditions for the state-space model are set to $\mathrm{E}[\mathbf{x}_0] = [100\ 10\ 30\ -10]^T$ and $\text{cov}[\mathbf{x}_0] = \text{diag}([25\ 2\ 25\ 2])$, where $\text{diag}(v)$ denotes a diagonal matrix with diagonal elements given by v.

In the first set of simulations, we assume that the diagonal element r in $\mathbf{R}$ is unknown and uniformly distributed over $[0.25, 4]$. We set q to 2. To analyze the average performances of different filtering approaches, for a fixed r, we calculate MSE for the OBKF, IBR, minimax, and MAP approaches as the trace of the estimation error covariance matrix obtained in (36) with $\mathbf{K}_k^{\boldsymbol{\theta}'}$ replaced by $\mathbf{K}_k^{\Theta^*}$ (obtained from (20)), $\mathbf{K}_k^{\Theta}$ (obtained relative to the prior distribution as in [9]), $\mathbf{K}_k^{\boldsymbol{\theta}_{\text{mm}}}$, and $\mathbf{K}_k^{\boldsymbol{\theta}_{\text{MAP}}}$, $\boldsymbol{\theta}_{\text{MAP}}$ being the MAP estimate, respectively. The MSE results shown in Fig. 3 are averaged over 30 different values of r generated

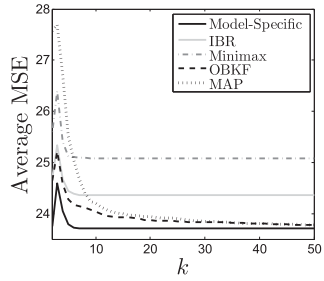

Fig. 3. Average MSE over time when r is unknown.

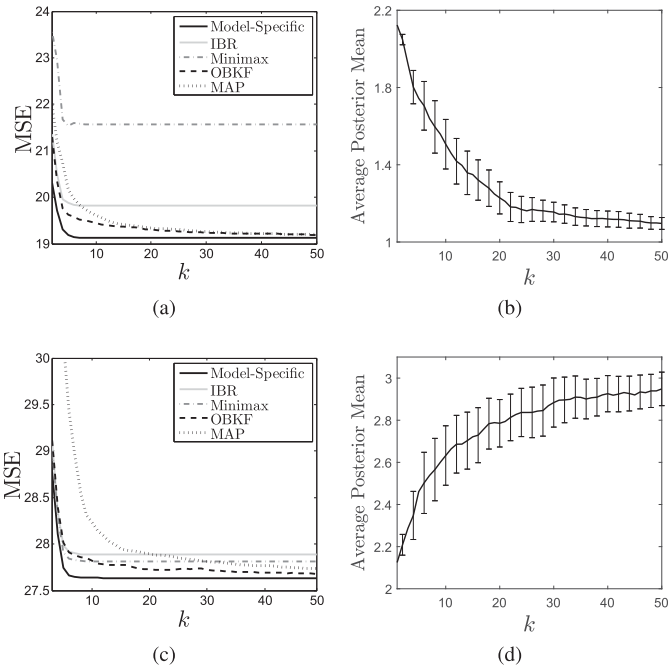

Fig. 4. Performance analysis for specific values of r. (a) MSE for $r = 1$. (b) Empirical average and variance of $\mathrm{E}[r|\mathcal{Y}_k]$ for $r = 1$. (c) MSE for $r = 3$. (d) Empirical average and variance of $\mathrm{E}[r|\mathcal{Y}_k]$ for $r = 3$.

according to the prior distribution with 10 different sets of observations for each. This figure suggests that in the beginning, the OBKF and IBR Kalman filters have almost the same average MSE but as more observations are incorporated the average MSE of the OBKF gets closer to that of the model-specific Kalman filter. Both the OBKF and IBR Kalman filters outperform the minimax Kalman in terms of average MSE. We can also see that in the beginning the OBKF approach outperforms the MAP approach but since both the posterior effective noise statistics and the MAP estimate converge to the true value when more observations are utilized, both approaches have similar performances for large k.

Fig. 4 compares different filters for two specific values $r = 1$ and $r = 3$. Throughout the paper whenever we focus on a specific model, the results are averaged over 200 different sets of observations generated according to the true model. In Fig. 4(a), the MSE of the OBKF goes from values close to the IBR to values close to the optimal model-specific Kalman filter. We also study the posterior mean computed for r using the approach outlined in Algorithms 2 and 3. Fig. 4(b) shows the empirical average and variance of the posterior mean for r obtained over different sets of observations. The variances are shown as vertical bars, where the length of each bar equals the variance. In the beginning, the posterior mean is close to the prior mean 2.125 but as more data are observed, the posterior distribution gets more concentrated around the true value and therefore the posterior mean tends toward $r = 1$. That is why the OBKF performs similarly as the IBR Kalman when few observations are utilized and has a close performance to the model-specific Kalman filter when more data are incorporated. Also, the variance of the posterior means gets smaller as we observe more data. One difference between Figs. 4(a) and (c) is that while the IBR Kalman filter outperforms the minimax Kalman filter in (a), the minimax Kalman filter has better performance in (b). Therefore, for the first few observations the OBKF has MSE larger than that of the minimax Kalman filter. Note that as discussed in [9], the IBR approach is optimal on average relative to the prior not for each specific model. Therefore, while it is guaranteed that the IBR approach results in the lower average MSE compared to the minimax approach, it might not outperform minimax for some specific models. For example, in Fig. 4(c) minimax outperforms IBR but as the OBKF is designed relative to the posterior distribution and the posterior effective noise statistics tend to the underlying true model as more observations are utilized, OBKF eventually outperforms (after first few observations) the minimax approach even for the models that the IBR approach fails to do so.

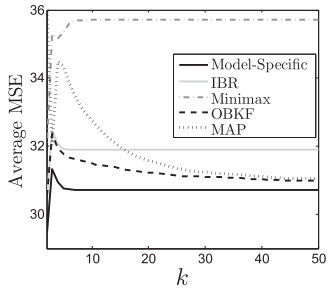

Fig. 5. Average MSE over time obtained by different Kalman filtering approaches when the process noise intensity q and the observation noise variance r are unknown.

In the second set of simulations, we assume that both the process noise intensity q and the observation noise diagonal element r are unknown. The uncertainty intervals are $q \in [1, 5]$ and $r \in [0.25, 4]$. The unknown parameters are independent and uniformly distributed. Fig. 5 presents the average MSE obtained for various filtering approaches averaged over 300 different combinations of q and r, and 10 different sets of observations for each combination. This figure demonstrates the promising performance of the OBKF compared to other robust filters.

In Fig. 6, we study how different filters perform for specific values of q and r. Each row corresponds to a specific pair of true values for q and r. The results are obtained based on different observation sequences generated according to the underlying true model. The MSE achieved by the OBKF gets closer to that of the optimal model-specific Kalman filter with more observations. Also, the average of the posterior means for q and r converge to the assumed true values when more observations are incorporated.

In Fig. 7, we study the effect of the tightness of the prior on the performance of different robust filtering strategies. Consider a Beta distribution $\mathcal{B}(\gamma\alpha, \gamma\beta)$ over $[0, 1]$, where $\alpha + \beta = 1$

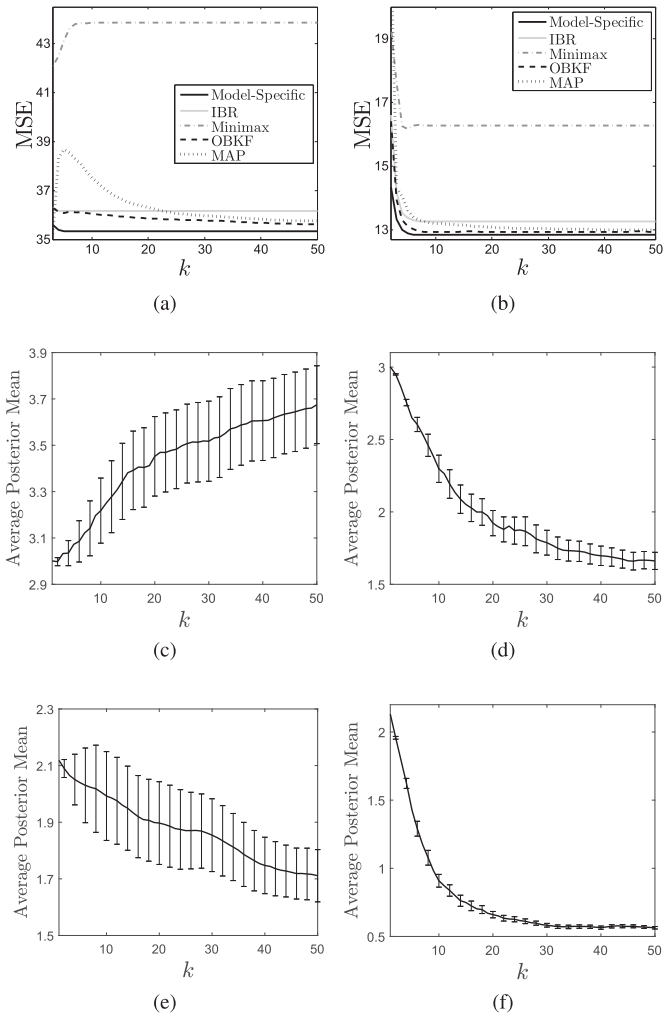

Fig. 6. Performance analysis for specific models. The first column corresponds to $q = 4$ and $r = 1.5$ and the second column corresponds to $q = 1.5$ and $r = 0.5$. Figures (a) and (b) show the MSE. Figures (c) and (d) show the empirical average and variance for $\mathrm{E}[q|\mathcal{Y}_k]$. Figures (e) and (f) show the empirical average and variance for $\mathrm{E}[r|\mathcal{Y}_k]$.

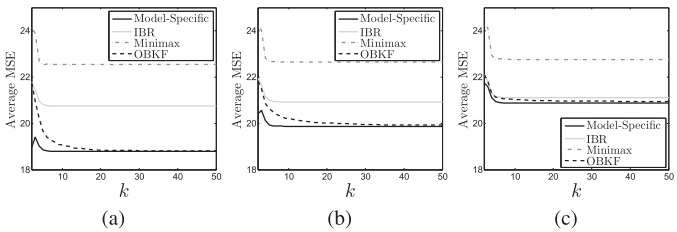

Fig. 7. Average performances of different filtering approaches when Beta priors with different γ are considered. (a) $\gamma = 0.1$. (b) $\gamma = 1$. (c) $\gamma = 10$.

and the mean and variance of the distribution are α and $\frac{\alpha\beta}{\gamma+1}$, respectively. Therefore, increasing γ, while the mean remains unchanged, leads to a smaller variance or tighter prior. In the simulations, we assume that $q = 2$ and r is unknown and governed by a scaled Beta distribution $\mathcal{B}(\gamma\alpha, \gamma\beta)$ over $[0.25, 4]$ with $\alpha = 0.3$ and $\beta = 0.7$. Fig. 7 presents the average MSE over different assumed true values for r generated according to the Beta priors with different γ. Since the mean of the prior is the same

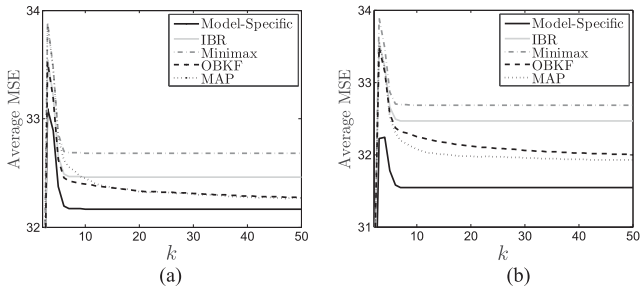

Fig. 8. Average performances of various filtering approaches when r is unknown and the prior distribution is falsely assumed as uniform distribution over interval $[3, 6]$. (a) True uncertainty interval is $[2, 7]$. (b) True uncertainty interval is $[0.5, 8.5]$.

for all distributions, the IBR Kalman filter has approximately the same average MSE for different values of γ. Also because minimax does not take into account the prior distribution, the minimax robust filter is the same for all three priors, resulting in the same average MSE performance. We can observe that as γ decreases (prior becomes looser), the difference between the average MSEs of the IBR and minimax filters and the average of the optimal MSEs obtained by optimal model-specific Kalman filters increases; however, average MSE for the OBKF converges to the average optimal MSE for all cases.

Next we study the effect of using an incorrect prior distribution for designing the OBKF. It is important for a robust filter to demonstrate some robustness relative to the inaccuracies in the prior distribution. To study the effect of using a bad prior distribution on the performance of a robust Kalman filter, we consider the case that the diagonal element r in the observation noise covariance matrix is unknown. Despite previous simulations where we have always assumed that the correct prior distribution is available, here we use a prior distribution that might not include the underlying true value for r. In Fig. 8, we design various robust Kalman filters when a uniform prior distribution over $[3, 6]$ is assumed for r. However, the correct prior distribution is different and is a uniform distribution over $[2, 7]$ and $[0.5, 8.5]$ in Figs. 8(a) and (b), respectively. We report the average MSE over time for various filters where the average is over 30 different true values based on the correct prior distribution and 20 different observation sequences for each true value. Despite in Fig. 3 where the OBKF and MAP-based approach eventually converge to the true Kalman filter designed relative to the true value, in Fig. 8 we observe that neither the OBKF nor MAP-based filter converges. This is because the true model might not belong to the assumed prior distribution used for designing the OBKF and as a result the posterior effective noise statistics do not converge to the underlying true value even for a large number of observations. However, the OBKF still demonstrates reasonable performance and outperforms other robust filters. Note that as the difference between the correct prior and the assumed distribution gets larger, the OBKF may perform poorer as its performance relies on the accuracy of the assumed prior distribution. For example, in Fig. 8(b) where the difference between the assumed and correct prior distributions is larger, we can see a larger gap between the OBKF performance and the optimal performance, and even the MAP-based approach is slightly better than that of the OBKF.

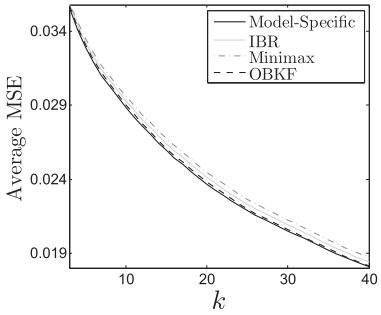

Fig. 9. Average MSE of different filtering approaches for the GRN inference when the observation noise parameter is unknown.

B. Example: Gene Regulatory Network Inference

In this section, we focus on the problem of gene regulatory network (GRN) inference using Kalman filtering. Due to the presence of noise in data acquisition and the complexity of the underlying biological system being modeled, robust inference is of high practical significance. We use the continuous nonlinear ordinary differential equation model for GRNs as in [35]. Suppose there are l genes g_i, $1 \leq i \leq l$, regulating each other such that the dynamics of the network are characterized by $\dot{g}_i = \eta_i(g_1, ..., g_l) + v_i$, where the dot notation $\dot{g}_i$ represents the derivative with respect to time variable t, v_i denotes the external noise, and the regulatory function η_i is of the form $\eta_i(g_1, ..., g_l) = \sum_{j=1}^{N_i} \left[(\gamma_{ij} + \epsilon_{ij})\Omega_{ij}(g_1, ..., g_l)\right]$, where γ_{ij} denotes the j-th coefficient in η_i corresponding to the parameter noise ϵ_{ij} and the nonlinear term $\Omega_{ij}(g_1, ..., g_l)$. The number of nonlinear terms in η_i is determined by N_i. We are interested in inferring the coefficients γ_{ij} via data generated from the network. This inference problem can be modeled as a state-space model in which $\mathbf{\Phi}_k = \mathbf{I}$ and $\mathbf{\Gamma}_k = \mathbf{I}$. Consider the yeast cell cycle network explained in [35] consisting of 12 genes and 54 coefficients γ_{ij}. For this network, the state vector $\mathbf{x}_k$ contains 54 unknown coefficients, the observation vector $\mathbf{y}_k$ is of size 12 and denotes the rate at which gene values change, and $\mathbf{u}_k$ and $\mathbf{v}_k$ represent the parameter noise ϵ_{ij} and the external noise $\mathbf{v}_i$, respectively. The initial conditions are set to $\mathrm{E}[\mathbf{x}_0] = \mathbf{0}_{54\times 1}$ and $\mathrm{cov}[\mathbf{x}_0] = 0.001 \times \mathbf{I}$. More details about the yeast cell cycle network and the procedure of deriving the state-space model for inference can be found in [9], [35].

We first analyze the performance of GRN inference for $\mathbf{Q} = 10^{-7} \times \mathbf{I}$ and $\mathbf{R} = r \times \mathbf{I}$, r being uniformly distributed over $[0.25, 6]$. In Fig. 9, we analyze the average MSE obtained via different filtering approaches. As can be seen, the average MSE of the OBKF is getting closer to that of the optimal model-specific Kalman filters as more data are observed. Also, note that both the IBR and OBKF outperform the minimax approach. Fig. 10 presents the results when $r = 1.5$ and $r = 5$. We can see that in both cases, OBKF always outperforms the IBR approach. Also, as the number of observations k increases, the empirical average of the posterior means computed for r goes towards the underlying true value. As mentioned earlier, since the IBR filter is optimal relative to the whole uncertainty class, it might not perform better than the minimax approach for some specific models inside the uncertainty class as is the case in Fig. 10(c) but for all states after some number of observations the OBKF approach outperforms the minimax strategy.

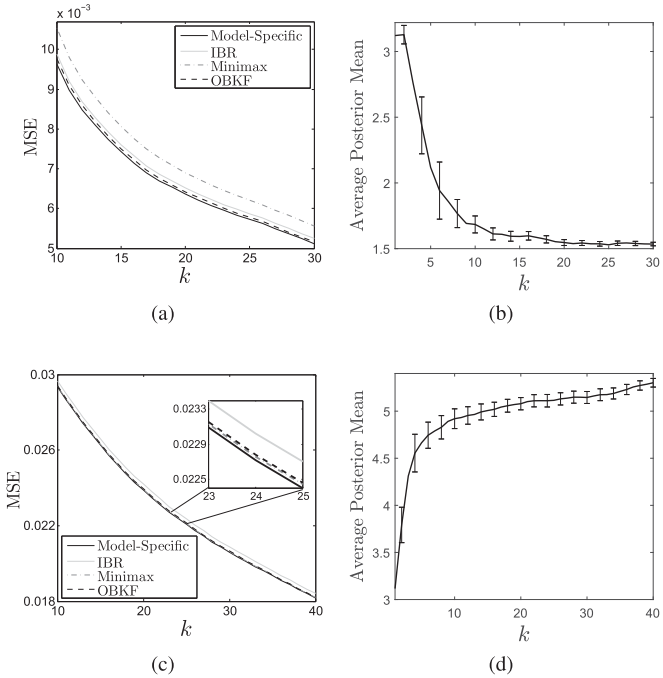

Fig. 10. Performance analysis for the GRN inference problem when r is unknown. (a) MSE for $r = 1.5$. (b) Empirical average and variance of posterior mean for $r = 1.5$. (c) MSE for $r = 5$. (d) Empirical average and variance of posterior mean for $r = 5$.

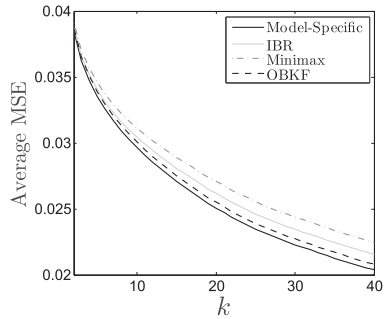

Fig. 11. Average MSE of different Kalman filtering approaches for GRN inference problem when both noise covariance matrices are unknown.

Now we assume that both the diagonal elements q and r of the process noise and observation noise covariance matrices are unknown. The uncertainty intervals for the unknown parameters are $q \in [10^{-8}, 10^{-6}]$ and $r \in [0.25, 6]$. Fig. 11 compares different filtering approaches in terms of the average MSE. As can be seen, the average MSE of the OBKF is always between those of the IBR and optimal model-specific filters.

In Fig. 12, we study the performance relative to the specific models inside the uncertainty class. Figures in the first column correspond to $q = 10^{-7}$ and $r = 5$ and those in the second column correspond to $q = 7 \times 10^{-7}$ and $r = 5.5$. OBKF-based inference always outperforms the IBR-based method, especially when enough observations are used for inference. Note that as discussed for Fig. 4, again we can see in Fig. 12(b) that while the minimax approach outperforms IBR (as the IBR optimization is relative to the prior not for each specific model), the OBKF still outperforms the minimax approach. We also show the empirical

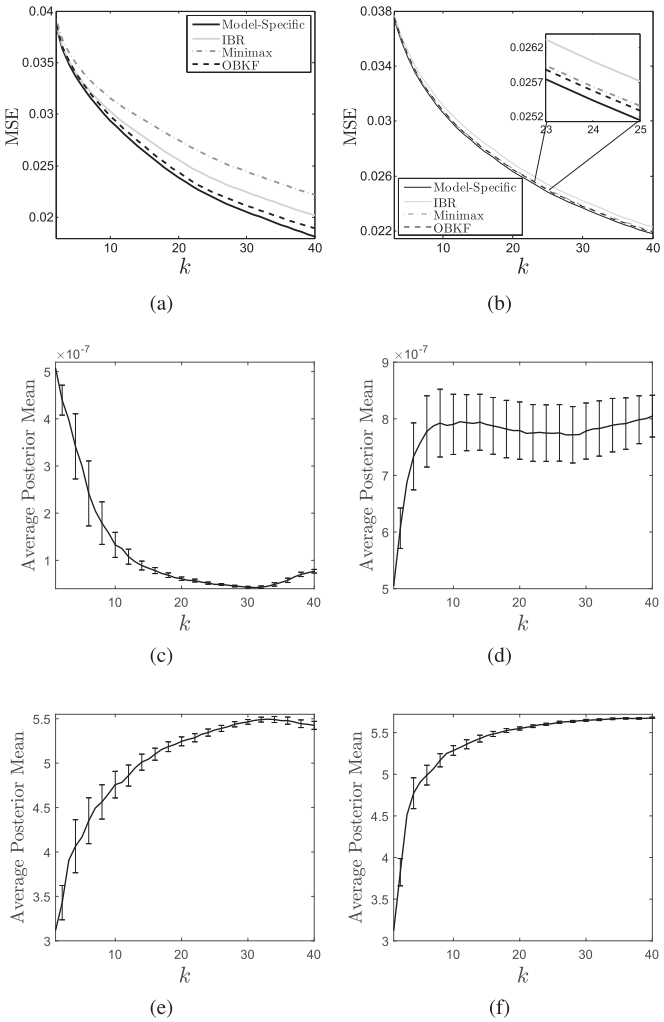

Fig. 12. Performance analysis for specific models. The first and second columns correspond to $q = 10^{-7}$, $r = 5$ and $q = 7 \times 10^{-7}$, $r = 5.5$, respectively. The first, second, and third rows present the MSE over time, the empirical average and variance for the posterior mean of q, and the empirical average and variance for the posterior mean of r, respectively.

average and variance for the posterior means of the unknown parameters. Note that to be able to plot the variance and average of the posterior mean for q in the same graph, we multiply the variance by 10^7. The empirical average for the posterior means computed using the factor graph step converge to the assumed true values.

C. Complexity Analysis

Regarding the computational complexity of the proposed OBKF framework, its recursive structure is completely similar to that of the ordinary Kalman filter except for the use of effective characteristics. Therefore, the main computational burden is in calculating the posterior effective noise statistics using the factor-graph-based approach. Assuming the aim is to calculate the posterior effective noise statistics at time k, for each generated MCMC sample, (27)–(30), should be computed from $i = 0$ to $i = k - 1$. Therefore, in addition to the dimension of the state vectors $\mathbf{x}$, the computational complexity also depends on both the number of MCMC samples and k. The complexity

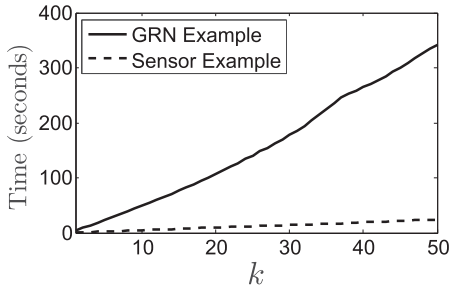

Fig. 13. Processing times required for calculating the posterior effective noise statistics for the sensor and GRN inference examples.

TABLE II
NUMBER OF FLOPS FOR EACH CALCULATION OF THE RECURSIVE STEP IN ALGORITHM 3

Line	Number of flops
Line 7	$3n^3 + 8.5n^2 + 6.5n + 3m^3 + 6.5m^2 + 7.5m + 2mn - 16$
Line 8	$10n^3 + 6n^2 + 17n + 2nm^2 + 2n^2m - nm - 16$
Line 9	$6n^3 - 2n^2$
Line 10	$4n^3 - 2n^2 + 2nm$
Line 11	$\frac{4}{3}n^3 + 20n^2 + \frac{68}{3}n + 85$

of computing (27) can be ignored compared to that of the other three equations. Moreover, all matrix inversions, matrix multiplications, and determinants might not need to be computed for each i and each MCMC sample. For example, when the process noise covariance matrix is known, $|\widetilde{\mathbf{Q}}_i|$ and $\widetilde{\mathbf{Q}}_i^{\theta_1}$ should be computed only one time, or when the system is stationary, $\mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}$ is independent from i and need be evaluated only one time. Fig. 13 shows average run times on a 2.9 GHz Intel Xeon, 8 GB RAM machine with codes written in MATLAB to compute the posterior effective noise statistics for the two applications studied in the paper when both observation and process noise parameters are unknown. The number of MCMC samples generated is 10,000. Since the dimension of the GRN and sensor examples are 54 and 4, respectively, the run times required for the former is much larger than that of the latter.

We also study computational complexity in terms of the number of real floating point operations (flops) [36]. We should note that although counting flops cannot provide exact computational complexity, it can reflect the order of complexity. We first provide the flop count of several basic scalar and matrix operations. Each real-valued scalar addition and multiplication requires one flop. Exponential function and taking the square root take 40 and 8 flops, respectively. The multiplication of two matrices of size $l \times n$ and $n \times m$ needs $l \times m \times (2n - 1)$ flops. Also, the inversion and determinant of an $n \times n$ matrix are counted as $3n^3 + 4.5n^2 + 8.5n - 8$ and $\frac{1}{3}n^3 + 4n^2 + \frac{14}{3}n$ flops, respectively. Using these flop counts, the number of flops for each line of the recursive step in Algorithm 3 is given in Table II. Note that the number of flops for each operation is counted only one time in this table. For example, the flops for the inversion of $\mathbf{\Sigma}_i$ or the multiplication $\mathbf{\Sigma}_i^{-1}\mathbf{M}_i$ are counted only for Line 7. Recall that n is the dimension of the state vector and m is the dimension of the observation vector in the state-space model.

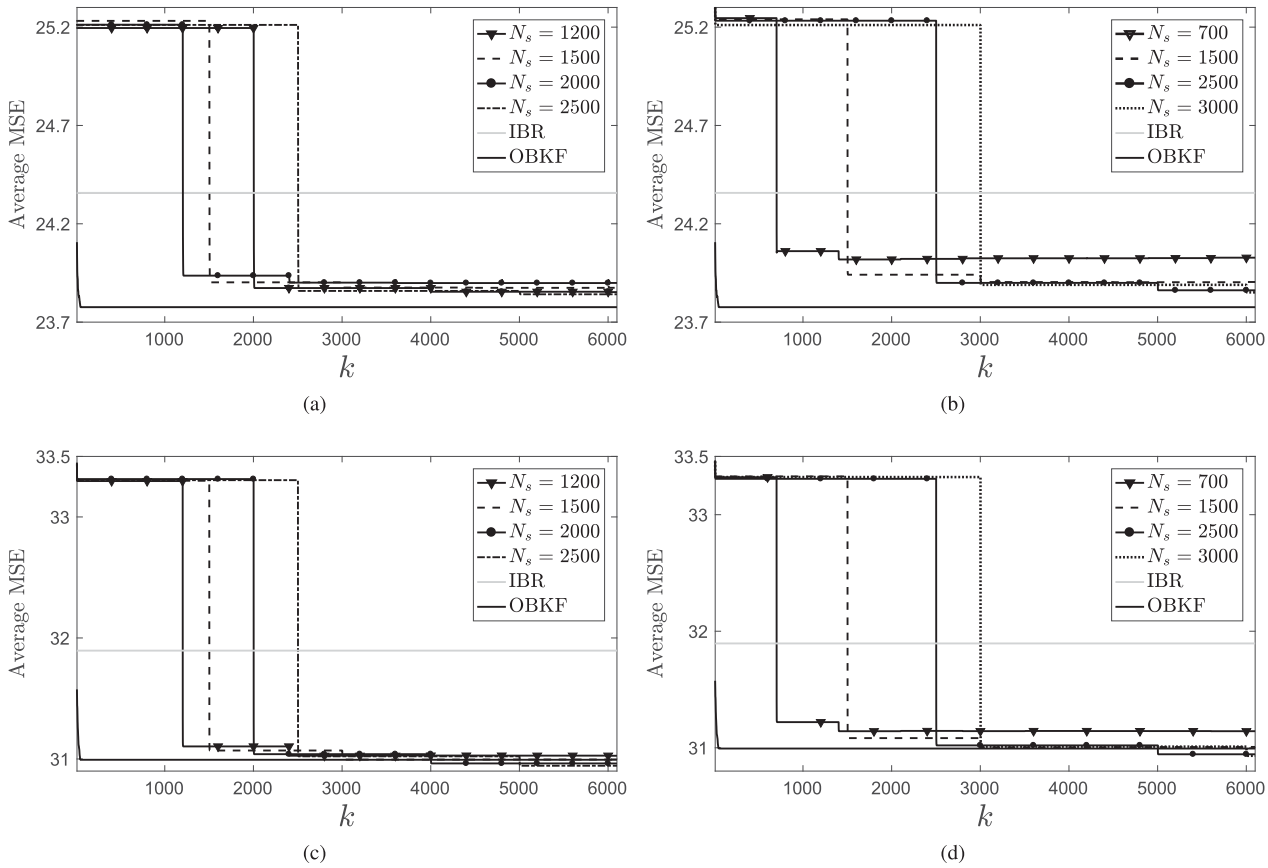

Fig. 14. Comparison with adaptive Kalman filters. (a) Unknown r and comparison with the Myers method. (b) Unknown r and comparison with the Mehra method. (c) Unknown q and r and comparison with the Myers method. (d) Unknown q and r and comparison with the Mehra method.

When n and m are large the order of the number of flops for lines 7, 8, 9, 10, and 11 are $\mathcal{O}(3n^3 + 3m^3)$, $\mathcal{O}(10n^3 + 2nm^2 + 2n^2m)$, $\mathcal{O}(6n^3)$, $\mathcal{O}(4n^3)$, and $\mathcal{O}(\frac{4}{3}n^3)$, respectively. Also, the order of flop count for each recursion in Algorithm 3 is $\mathcal{O}(24n^3 + 3m^3 + 2nm^2 + 2n^2m)$. This order of flop count reflects the order of complexity for each recursion from $i = 0$ to $i = k - 1$, k being the time index of estimation, and for each generated MCMC sample.

D. Comparison With the Adaptive Kalman Filtering Approach

Adaptive Kalman methods aim to simultaneously estimate both the states and the noise statistics [14]–[17]. In this section, we compare the OBKF with two benchmark adaptive Kalman methods proposed by Myers [15] and Mehra [16].

The Myers method involves the empirical estimation of the noise statistics based on observations. For each batch of N_s observations, first the observation and process noise samples are approximated using the estimated states and then the noise samples are used for unbiased noise covariance estimation. This process of updating noise covariances can be repeated after every N_s observations.

In the Mehra method [16], rather than first estimating the unknown noise covariance, the aim is to use the innovation process to directly estimate the steady-state value of the Kalman gain matrix. This method uses the fact that when the Kalman filter is designed relative to the true noise statistics, the resulting innovation would be a white-noise process but when noise covariance matrices are unknown, the autocorrelation function of the innovation process, which is not zero anymore, can be used to estimate the Kalman gain matrix. After observing every N_s samples, the sample autocorrelation function for the innovation process is computed and then used to estimate the Kalman gain matrix. To apply this method, the system being studied has to be time-invariant and completely controllable and observable.

Fig. 14 compares the OBKF framework with the two aforementioned adaptive filtering approaches for the target tracking example in Section IV-A. The performances of the adaptive Kalman filters depend on the number of observations N_s used for tuning the filter. If a small N_s is chosen, it is likely that highly inaccurate adjustment of the filter parameters after each iteration causes larger state estimation errors for future observations, which in turn may lead to a poorer adjustment for the next round, which in the long run may eventually make the filter unstable. On the other hand, a large N_s would delay the adjustment process, thereby resulting in a large number of states being estimated using old unadjusted parameters. We consider different N_s for the adaptive filtering methods. In Figs. 14(a) and (b), we assume that the diagonal element r of the observation noise covariance matrix is uniformly distributed over $[0.25, 4]$. Figs. 14(c) and (d) study the case that in addition to r, the intensity q of the process noise is unknown and uniformly distributed over $[1, 5]$. To obtain the average MSE for the adaptive methods, we run 10,000 simulations over different assumed true values for the unknown parameters and different sets of observations for each. The average MSEs for adaptive filters are obtained by

taking the average of the MSEs, calculated according to (36), for different simulations. The average MSEs of the IBR and OBKF methods seem to be constant. This is because the IBR approach converges quickly to its steady-state value, for the OBKF, we calculate the posterior effective noise statistics for up to $k = 50$ observations and after that we use the last obtained posterior effective characteristics to find the state estimates for the remaining observations. When r is unknown ((a) and (b)), the OBKF approach yields lower average MSE compared to the adaptive methods for each number of observations. When both r and q are unknown ((c) and (d)), the performance of the OBKF is much better than those of the adaptive methods when k is small and is comparable when k is large. Note that in many practical applications it is not realistic to assume a generous amount of data ($k = 6100$). For example, in the case of Kalman-based GRN identification, genomic data are expensive and usually obtained from living organisms. Thus, inference must be done with a limited amount of data. Therefore, the performance analysis in the case of small samples is of high practical significance.

V. Conclusion

This paper extends the theory of intrinsically Bayesian robust Kalman filtering to optimal Bayesian Kalman filtering by performing the optimization relative to the posterior distribution of the unknown noise parameters. The optimal Bayesian Kalman filter is designed by introducing the concept of posterior effective noise statistics that can be computed using the proposed factor-graph-based algorithm.

One issue of the optimal Bayesian Kalman filtering is that the factor-graph-based step might be computationally expensive when a large number of observations are considered. Future work includes finding more efficient computation approaches for the MCMC step. We should emphasize that when the posterior effective noise statistics are computed for a reasonable amount of data, the MCMC step can be dropped from filtering and the last computed posterior effective noise statistics can be used for the rest of observations.

Another important future work that has significant practical implications is designing experiments for optimally reducing the uncertainty in the state-space model. The proposed OBKF framework enables us to utilize the concept of Mean Objective Cost of Uncertainty (MOCU) [27] for quantifying uncertainty in an objective-based manner relative to the performance of the robust operator in the presence of uncertainty, herein, the optimal Bayesian Kalman filter. MOCU-based experimental design can lead us to the experiment with the most potential for reducing the pertinent uncertainty in the model [37], [38].

Finally, it is worth noting that while the Bayesian method proposed in this paper is the optimal solution to the Kalman filter problem in the presence of uncertainty in noise statistics, the entire framework depends on the prior probabilities and their relevance to the user's knowledge regarding the underlying system. A body of work has been recently devoted to the construction of such priors in gene regulatory networks [39], [40], and we plan to take such an approach to construct priors for the OBKF. In particular, we envision that by combining some initial runs (i.e. time series) of the dynamical system with the literature-driven knowledge, e.g. gene regulatory networks or the noise characteristics, commonly known in different technology-specific data generating methods, we can objectively build informative priors. As an example, one can take existing gene expression data from next-generation sequencing platforms, quantify the noise, build the corresponding noise statistics priors, and simultaneously constrain the state update model with known gene-gene interactions to come up with a unified model that is both knowledge- and data-driven. This is practically important for any application with small sample data sets, including genomics and geoscience.

Appendix
Proof of Lemma 2

Since $\mathbf{x}_i$ (and $\mathbf{x}_{i+1}$) is of size $n \times 1$ and $\mathbf{y}_i$ is of size $m \times 1$, we rewrite the integrand of (26) as:

$$\frac{S_i}{\sqrt{(2\pi)^n \left|\widetilde{\mathbf{Q}}_i^{\theta_1}\right|}\sqrt{(2\pi)^m \left|\mathbf{R}^{\theta_2}\right|}\sqrt{(2\pi)^n |\mathbf{\Sigma}_i|}} \times \exp\left(\frac{-1}{2}\left(\left(\mathbf{x}_{i+1} - \mathbf{\Phi}_i\mathbf{x}_i\right)^T (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\left(\mathbf{x}_{i+1} - \mathbf{\Phi}_i\mathbf{x}_i\right) + \left(\mathbf{y}_i - \mathbf{H}_i\mathbf{x}_i\right)^T (\mathbf{R}^{\theta_2})^{-1}\left(\mathbf{y}_i - \mathbf{H}_i\mathbf{x}_i\right) + \left(\mathbf{x}_i - \mathbf{M}_i\right)^T \mathbf{\Sigma}_i^{-1}\left(\mathbf{x}_i - \mathbf{M}_i\right)\right)\right). \tag{37}$$

The argument of the exponential function can be expanded as

$$\begin{aligned}&\mathbf{x}_{i+1}^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{x}_{i+1} - \mathbf{x}_{i+1}^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{\Phi}_i\mathbf{x}_i - \mathbf{x}_i^T\mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{x}_{i+1}\\ &+ \mathbf{x}_i^T\mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{\Phi}_i\mathbf{x}_i + \mathbf{y}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i - \mathbf{x}_i^T\mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i\\ &- \mathbf{y}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_i\mathbf{x}_i + \mathbf{x}_i^T\mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_i\mathbf{x}_i + \mathbf{x}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{x}_i\\ &- \mathbf{M}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{x}_i - \mathbf{x}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{M}_i + \mathbf{M}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{M}_i.\end{aligned} \tag{38}$$

We want to complete the square for $\mathbf{x}_i$. Thus, we rearrange the terms in (38) involving $\mathbf{x}_i$ as

$$\begin{aligned}&\mathbf{x}_i^T\left(\mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{\Phi}_i + \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_i + \mathbf{\Sigma}_i^{-1}\right)\mathbf{x}_i\\ &+ \mathbf{x}_i^T\left(-\mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{x}_{i+1} - \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i - \mathbf{\Sigma}_i^{-1}\mathbf{M}_i\right)\\ &+ \left(-\mathbf{x}_{i+1}^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{\Phi}_i - \mathbf{y}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_i - \mathbf{M}_i^T\mathbf{\Sigma}_i^{-1}\right)\mathbf{x}_i\\ &+ \mathbf{y}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i + \mathbf{M}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{M}_i.\end{aligned} \tag{39}$$

Plugging

$$\mathbf{\Lambda}_i^{-1} = \mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{\Phi}_i + \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{H}_i + \mathbf{\Sigma}_i^{-1}, \tag{40}$$

$$\mathbf{G}_i = \mathbf{\Lambda}_i\left(\mathbf{\Phi}_i^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{x}_{i+1} + \mathbf{H}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i + \mathbf{\Sigma}_i^{-1}\mathbf{M}_i\right), \tag{41}$$

in (39) yields

$$\left(\mathbf{x}_i - \mathbf{G}_i\right)^T\mathbf{\Lambda}_i^{-1}\left(\mathbf{x}_i - \mathbf{G}_i\right) - \mathbf{G}_i^T\mathbf{\Lambda}_i^{-1}\mathbf{G}_i + \mathbf{x}_{i+1}^T(\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}\mathbf{x}_{i+1} + \mathbf{y}_i^T(\mathbf{R}^{\theta_2})^{-1}\mathbf{y}_i + \mathbf{M}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{M}_i. \tag{42}$$

We can further simplify (42). Substituting (41) for $\mathbf{G}_i$ in the term $\mathbf{G}_i^T \boldsymbol{\Lambda}_i^{-1} \mathbf{G}_i$ in (42) and letting

$$\boldsymbol{\Sigma}_{i+1}^{-1} = (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1} - (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1} \boldsymbol{\Phi}_i \boldsymbol{\Lambda}_i \boldsymbol{\Phi}_i^T (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1}, \tag{43}$$

$$\mathbf{M}_{i+1} = \boldsymbol{\Sigma}_{i+1} (\widetilde{\mathbf{Q}}_i^{\theta_1})^{-1} \boldsymbol{\Phi}_i \boldsymbol{\Lambda}_i \Big(\mathbf{H}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_i + \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i \Big), \tag{44}$$

we can rewrite (42) as

$$\begin{aligned}
&\big(\mathbf{x}_i - \mathbf{G}_i\big)^T \boldsymbol{\Lambda}_i^{-1} \big(\mathbf{x}_i - \mathbf{G}_i\big) + \big(\mathbf{x}_{i+1} - \mathbf{M}_{i+1}\big)^T \\
&\times \ \boldsymbol{\Sigma}_{i+1}^{-1} \big(\mathbf{x}_{i+1} - \mathbf{M}_{i+1}\big) \\
&- \ \mathbf{M}_{i+1}^T \boldsymbol{\Sigma}_{i+1}^{-1} \mathbf{M}_{i+1} - \mathbf{y}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{H}_i \boldsymbol{\Lambda}_i \mathbf{H}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_i \\
&- \ \mathbf{y}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{H}_i \boldsymbol{\Lambda}_i \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i - \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Lambda}_i \mathbf{H}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_i \\
&+ \ \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Lambda}_i \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i + \mathbf{y}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_i + \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i \\
&= \big(\mathbf{x}_i - \mathbf{G}_i\big)^T \boldsymbol{\Lambda}_i^{-1} \big(\mathbf{x}_i - \mathbf{G}_i\big) \\
&+ \ \big(\mathbf{x}_{i+1} - \mathbf{M}_{i+1}\big)^T \boldsymbol{\Sigma}_{i+1}^{-1} \big(\mathbf{x}_{i+1} - \mathbf{M}_{i+1}\big) - \mathbf{M}_{i+1}^T \boldsymbol{\Sigma}_{i+1}^{-1} \mathbf{M}_{i+1} \\
&- \ \mathbf{W}_i^T \boldsymbol{\Lambda}_i \mathbf{W}_i + \mathbf{y}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_i + \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i,
\end{aligned} \tag{45}$$

where $\mathbf{W}_i = \mathbf{H}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_i + \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i$. Substituting (45) as the argument of the exponential function in (37) gives

$$\begin{aligned}
&\frac{S_i}{\sqrt{(2\pi)^n \big|\widetilde{\mathbf{Q}}_i^{\theta_1}\big|} \sqrt{(2\pi)^m \big|\mathbf{R}^{\theta_2}\big|} \sqrt{(2\pi)^n |\boldsymbol{\Sigma}_i|}} \\
&\times \ \exp\Bigg(\frac{-1}{2} \Bigg(\big(\mathbf{x}_i - \mathbf{G}_i\big)^T \boldsymbol{\Lambda}_i^{-1} \big(\mathbf{x}_i - \mathbf{G}_i\big) \\
&+ \ \big(\mathbf{x}_{i+1} - \mathbf{M}_{i+1}\big)^T \boldsymbol{\Sigma}_{i+1}^{-1} \big(\mathbf{x}_{i+1} - \mathbf{M}_{i+1}\big) - \mathbf{M}_{i+1}^T \boldsymbol{\Sigma}_{i+1}^{-1} \mathbf{M}_{i+1} \\
&\qquad - \ \mathbf{W}_i^T \boldsymbol{\Lambda}_i \mathbf{W}_i + \mathbf{y}_i^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_i + \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i \Bigg) \Bigg) \\
&= S_i \frac{\sqrt{|\boldsymbol{\Lambda}_i|}}{\sqrt{|\widetilde{\mathbf{Q}}_i^{\theta_1}|}} \mathcal{N}(\mathbf{x}_i; \mathbf{G}_i, \boldsymbol{\Lambda}_i) \frac{\sqrt{|\boldsymbol{\Sigma}_{i+1}|}}{\sqrt{|\boldsymbol{\Sigma}_i|}} \mathcal{N}(\mathbf{x}_{i+1}; \mathbf{M}_{i+1}, \boldsymbol{\Sigma}_{i+1}) \\
&\times \ \mathcal{N}\big(\mathbf{y}_i; \mathbf{0}_{m\times 1}, \mathbf{R}^{\theta_2}\big) \\
&\times \ \exp\Big(\frac{\mathbf{M}_{i+1}^T \boldsymbol{\Sigma}_{i+1}^{-1} \mathbf{M}_{i+1} + \mathbf{W}_i^T \boldsymbol{\Lambda}_i \mathbf{W}_i - \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i}{2} \Big).
\end{aligned} \tag{46}$$

We compute the original integral in (26) by integrating (46) relative to $\mathbf{x}_i$ that results in:

$$\begin{aligned}
&S_i \frac{\sqrt{|\boldsymbol{\Lambda}_i| \, |\boldsymbol{\Sigma}_{i+1}|}}{\sqrt{|\widetilde{\mathbf{Q}}_i^{\theta_1}| \, |\boldsymbol{\Sigma}_i|}} \mathcal{N}(\mathbf{x}_{i+1}; \mathbf{M}_{i+1}, \boldsymbol{\Sigma}_{i+1}) \mathcal{N}\big(\mathbf{y}_i; \mathbf{0}_{m\times 1}, \mathbf{R}^{\theta_2}\big) \\
&\times \ \exp\Big(\frac{\mathbf{M}_{i+1}^T \boldsymbol{\Sigma}_{i+1}^{-1} \mathbf{M}_{i+1} + \mathbf{W}_i^T \boldsymbol{\Lambda}_i \mathbf{W}_i - \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i}{2} \Big).
\end{aligned} \tag{47}$$

Putting (47) in the form of $S_{i+1} \mathcal{N}(\mathbf{x}_{i+1}; \mathbf{M}_{i+1}, \boldsymbol{\Sigma}_{i+1})$, it can be seen that $\boldsymbol{\Sigma}_{i+1}$, $\mathbf{M}_{i+1}$, and S_{i+1} can be computed according to (27), (28), and (29), respectively.

REFERENCES

[1] V. P. Kuznetsov, "Stable detection when the signal and spectrum of normal noise are inaccurately known," *Telecommun. Radio Eng.*, vol. 3031, pp. 58–64, 1976.

[2] S. A. Kassam and T. L. Lim, "Robust Wiener filters," *J. Franklin Inst.*, vol. 304, no. 4, pp. 171–185, 1977.

[3] H. V. Poor, "Robust matched filters," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 677–687, Sep. 1983.

[4] S. Verdu and H. Poor, "Minimax linear observers and regulators for stochastic systems with uncertain second-order statistics," *IEEE Trans. Autom. Control*, vol. AC-29, no. 6, pp. 499–511, Jun. 1984.

[5] V. Poor and D. P. Looze, "Minimax state estimation for linear stochastic systems with noise uncertainty," *IEEE Trans. Autom. Control*, vol. AC-26, no. 4, pp. 902–906, Aug. 1981.

[6] A. M. Grigoryan and E. R. Dougherty, "Design and analysis of robust binary filters in the context of a prior distribution for the states of nature," *J. Math. Imag. Vis.*, vol. 11, no. 3, pp. 239–254, 1999.

[7] A. M. Grigoryan and E. R. Dougherty, "Bayesian robust optimal linear filters," *Signal Process.*, vol. 81, no. 12, pp. 2503–2521, 2001.

[8] L. Dalton and E. Dougherty, "Intrinsically optimal Bayesian robust filtering," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 657–670, Feb. 2014.

[9] R. Dehghannasiri, M. S. Esfahani, and E. R. Dougherty, "Intrinsically Bayesian robust Kalman filter: An innovation process approach," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2531–2546, May 2017.

[10] R. Dehghannasiri, X. Qian, and E. R. Dougherty, "Intrinsically Bayesian robust Karhunen-Loève compression," *Signal Process.*, vol. 144, pp. 311–322, 2018.

[11] R. Dehghannasiri, X. Qian, and E. R. Dougherty, "Optimal experimental design in the context of canonical expansions," *IET Signal Process.*, vol. 11, no. 8, pp. 942–951, 2017.

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

[13] S. Sangsuk-Iam and T. E. Bullock, "Analysis of discrete-time Kalman filtering under incorrect noise covariances," *IEEE Trans. Autom. Control*, vol. 35, no. 12, pp. 1304–1309, Dec. 1990.

[14] S. Särkkä and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *IEEE Trans. Autom. Control*, vol. 54, no. 3, pp. 596–600, Mar. 2009.

[15] K. A. Myers and B. D. Tapley, "Adaptive sequential estimation with unknown noise statistics," *IEEE Trans. Autom. Control*, vol. AC-21, no. 4, pp. 520–523, Aug. 1976.

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

[17] R. Mehra, "Approaches to adaptive filtering," *IEEE Trans. Autom. Control*, vol. AC-17, no. 5, pp. 693–698, Oct. 1972.

[18] J. M. Morris, "The Kalman filter: A robust estimator for some classes of linear quadratic problems," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 5, pp. 526–534, Sep. 1976.

[19] Y. S. Shmaliy, "An iterative Kalman-like algorithm ignoring noise and initial conditions," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2465–2473, Jun. 2011.

[20] W. H. Kwon, K. S. Lee, and O. K. Kwon, "Optimal FIR filters for time-varying state-space models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, no. 6, pp. 1011–1021, Nov. 1990.

[21] Y. S. Shmaliy, "Linear optimal FIR estimation of discrete time-invariant state-space models," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3086–3096, Jun. 2010.

[22] H. W. Bode and C. E. Shannon, "A simplified derivation of linear least square smoothing and prediction theory," *Proc. IRE*, vol. 38, no. 4, pp. 417–425, Apr. 1950.

[23] L. A. Zadeh and J. R. Ragazzini, "An extension of Wiener's theory of prediction," *J. Appl. Phys.*, vol. 21, no. 7, pp. 645–655, 1950.

[24] T. Kailath, "An innovations approach to least-squares estimation—Part I: Linear filtering in additive white noise," *IEEE Trans. Autom. Control*, vol. AC-13, no. 6, pp. 646–655, Dec. 1968.

[25] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[26] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Process. Mag.*, vol. 21, no. 1, pp. 28–41, Jan. 2004.

[27] B.-J. Yoon, X. Qian, and E. Dougherty, "Quantifying the objective cost of uncertainty in complex dynamical systems," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2256–2266, May 2013.

[28] L. A. Dalton and E. R. Dougherty, "Optimal classifiers with minimum expected error within a Bayesian framework—Part I: Discrete and Gaussian models," *Pattern Recognit.*, vol. 46, no. 5, pp. 1301–1314, 2013.
[29] X. Qian and E. Dougherty, "Bayesian regression with network prior: Optimal Bayesian filtering perspective," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6243–6253, Dec. 2016.
[30] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *J. Basic Eng.*, vol. 83, no. 1, pp. 95–108, 1961.
[31] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.
[32] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman & Hall/CRC Interdisciplinary Statistics). New York, NY, USA: Taylor & Francis, 1995.
[33] S. Challa, M. R. Morelande, D. Musicki, and R. J. Evans, *Fundamentals of Object Tracking*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[34] J. L. Williams, "Marginal multi-Bernoulli filters: RFS derivation of MHT, JIPDA, and association-based member," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 3, pp. 1664–1687, Jul. 2015.
[35] L. Qian, H. Wang, and E. R. Dougherty, "Inference of noisy nonlinear differential equation models for gene regulatory networks using genetic programming and Kalman filtering," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3327–3339, Jul. 2008.
[36] G. H. Golub and C. F. van Loan, *Matrix Computations*. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.
[37] R. Dehghannasiri, B.-J. Yoon, and E. Dougherty, "Optimal experimental design for gene regulatory networks in the presence of uncertainty," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 4, pp. 938–950, Jul. 2015.
[38] R. Dehghannasiri, B.-J. Yoon, and E. R. Dougherty, "Efficient experimental design for uncertainty reduction in gene regulatory networks," *BMC Bioinf.*, vol. 16, no. Suppl. 13, pp. 1–18, 2015.
[39] M. S. Esfahani and E. R. Dougherty, "Incorporation of biological pathway knowledge in the construction of priors for optimal Bayesian classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 1, pp. 202–218, Jan./Feb. 2014.
[40] M. S. Esfahani and E. R. Dougherty, "An optimization-based framework for the transformation of incomplete biological knowledge into a probabilistic structure and its application to the utilization of gene/protein signaling pathways in discrete phenotype classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 6, pp. 1304–1321, Nov./Dec. 2015.

Roozbeh Dehghannasiri (S'13) received the B.S. degree from the University of Tehran, Tehran, Iran, in 2010, the M.A.Sc. degree from the McMaster University, Hamilton, ON, Canada, in 2012, and the Ph.D. degree from Texas A&M University, College Station, TX, USA, in 2016, all in electrical engineering. Since the Ph.D. degree, he has been a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering, Texas A&M University. His research interests include statistical signal processing, uncertainty quantification in dynamical systems, robust filtering, and computational biology. He was the recipient of the Best Paper Award at the 12th Annual MCBIOS Conference in 2015 and the McMaster Outstanding Thesis Research Award in 2012.

Mohammad Shahrokh Esfahani (S'07–M'15) received the B.S. degree from the University of Tehran, Tehran, Iran, in 2007, the M.Sc. degree from Sharif University of Technology, Tehran, Iran, in 2009, and the Ph.D. degree from Texas A&M University, College Station, TX, USA, in 2014, all in electrical engineering. He is currently a Postdoctoral Research Fellow with the Division of Oncology and Center for Cancer Systems Biology, Stanford School of Medicine, Stanford University, Stanford, CA, USA. Prior to that, he was a Postdoctoral Research Associate with the Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University. His research interests include small-sample classification, Bayesian statistics, and statistical signal processing.

Xiaoning Qian (S'01–M'07–SM'17) received the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA, in 2005. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is affiliated with the Center for Bioinformatics and Genomic Systems Engineering and the Center for Translational Environmental Health Research, Texas A&M University. He is also a Courtesy Assistant Professor with the Department of Computer Science and Engineering and the Department of Pediatrics, University of South Florida, Tampa, FL, USA. His research interests include computational network biology, genomic signal processing, and biomedical image analysis.

Edward R. Dougherty (M'05–SM'09–F'12) received the Ph.D. degree in mathematics from Rutgers University, New Brunswick, NJ, USA, and the M.S. degree in computer science from Stevens Institute of Technology, Hoboken, NJ, USA. He is currently a Distinguished Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, where he holds the Robert M. Kennedy '26 Chair in electrical engineering and is the Scientific Director of the Center for Bioinformatics and Genomic Systems Engineering. He is the author of 16 books and the author of more than 300 journal papers. He was the recipient of the *Doctor Honoris Causa* by the Tampere University of Technology, and the SPIE President's Award. He is a Fellow of SPIE, and was the Editor of the SPIE/IS&T *Journal of Electronic Imaging*. At Texas A&M University, he was the recipient of the Association of Former Students Distinguished Achievement Award in Research, and was named Fellow of the Texas Engineering Experiment Station and Halliburton Professor of the Dwight Look College of Engineering.