

RESEARCH

Open Access



A Bayesian robust Kalman smoothing framework for state-space models with uncertain noise statistics

Roozbeh Dehghannasiri* , Xiaoning Qian and Edward R. Dougherty

Abstract

The classical Kalman smoother recursively estimates states over a finite time window using all observations in the window. In this paper, we assume that the parameters characterizing the second-order statistics of process and observation noise are unknown and propose an optimal Bayesian Kalman smoother (OBKS) to obtain smoothed estimates that are optimal relative to the posterior distribution of the unknown noise parameters. The method uses a Bayesian innovation process and a posterior-based Bayesian orthogonality principle. The optimal Bayesian Kalman smoother possesses the same forward-backward structure as that of the ordinary Kalman smoother with the ordinary noise statistics replaced by their effective counterparts. In the first step, the posterior effective noise statistics are computed. Then, using the obtained effective noise statistics, the optimal Bayesian Kalman filter is run in the forward direction over the window of observations. The Bayesian smoothed estimates are obtained in the backward step. We validate the performance of the proposed robust smoother in the target tracking and gene regulatory network inference problems.

Keywords: Kalman smoother, Robust filtering, Bayesian robustness, Innovation process, Orthogonality principle

1 Introduction

Classical Kalman filtering is defined via a set of equations that provide a recursive evaluation of the optimal linear filter output to incorporate new observations [1]. The filtering procedure assumes a state-space model consisting of a transition equation and an observation equation. There are three filtering paradigms [2]: the Kalman filter estimates the signal at the most recent observed time point, the Kalman predictor estimates the signal at a future time point, and the Kalman smoother estimates the signal at an intermediate observation time point. The equations for the filter and predictor are closely related, so that solving one provides an immediate solution for the other, whereas the smoother requires further work.

The issue that concerns us here is how to proceed when the model is not fully known. Classically, a precondition for optimal filtering is to have complete knowledge of the random process model; however, this assumption

is not always realistic in many practical settings such as target tracking [3–7]. Over the years, various adaptive procedures have been developed that essentially provide improving model estimates with increasing numbers of observations [8, 9]. More recently, the problem has been addressed under the assumption that the model belongs to an uncertainty class of models governed by a prior probability distribution, thereby placing the matter in a Bayesian framework with the aim being to find a recursive filter that is optimal over the uncertainty class.

There are two existing viewpoints for designing robust filters: minimax robustness, which involves designing a filter with the best worst-case performance [10–12], and Bayesian robustness, which involves designing a robust filter with the optimal performance on average relative to a prior (or posterior) distribution governing the uncertainty class [13–16]. When designing a Bayesian robust filter, if optimization is not constrained, meaning that it is over the entire class of filters of a particular type, then the filter is called an *intrinsically Bayesian robust filter* when optimality is relative to the prior distribution, and called an *optimal Bayesian filter* when optimality is relative to

*Correspondence: roozbehdn@email.tamu.edu
Department of Electrical and Computer Engineering, Texas A&M University,
College Station, Texas 77843, USA

the posterior distribution. This kind of uncertainty modeling has been applied to linear and morphological filtering, both with and without incorporating the information embedded in the observations into the prior distribution [15, 17]. In the case of Kalman filtering, the problem has been addressed for the filter and predictor without prior updating [16], which is called an *intrinsically Bayesian robust Kalman filter*, and with updating based on new data [18], which is called an *optimal Bayesian Kalman filter*. In this paper, we find the optimal Kalman smoother relative to the probability mass governing the uncertainty.

In a Bayesian robustness setting, the prior (posterior) distribution is on the model of the underlying random process, meaning that it refers directly to our scientific uncertainty. The general aim is to find an operator that is optimal with respect to both the stochasticity in the nominal problem, for which the underlying model is fully known, and the model uncertainty. The aim can be achieved by replacing model characteristics and statistics in the solution to the nominal problem with their *effective* counterparts, which incorporate model uncertainty in such a way that the equation structure of the nominal solution is essentially preserved in the Bayesian robust solution. This approach has been used for classification [19], linear and morphological filtering [15, 17], signal compression [20], and Kalman filtering [16]. For example, in optimal wide-sense stationary linear filtering, the power spectra are replaced by effective power spectra [15] or in Gaussian classification, the class-conditional densities are replaced by effective class-conditional densities [19].

An *intrinsically Bayesian robust Kalman filter* (IBR-KF) has been proposed in [16] that is optimal relative to the prior distribution of noise parameters. The theory of the IBR-KF is rooted in the Bayesian orthogonality principle and the Bayesian innovation process, which are the extended versions of their ordinary counterparts when applied to the prior distribution. Innovation processes have long been used for Kalman filtering, dating back to 1968 when Kailath proposed the first instance of an innovation-based approach for Kalman filtering [21]. Building on the IBR-KF theory developed in [16], an *optimal Bayesian Kalman filter* (OBKF) achieving optimality on average relative to the posterior distribution of the noise parameters when observations are incorporated into the prior distribution was proposed [18]. The OBKF shares the theoretical foundation of the IBR-KF, the difference being the distribution relative to which the Bayesian innovation process and Bayesian orthogonality principle are stated. It is the prior distribution in the latter [16] and the posterior distribution in the former [18].

Kalman smoothing is an offline signal processing tool where both past and future observations are used for making estimations [22–29]. Kalman smoothers can be

classified as fixed-point, fixed-lag, and fixed-interval smoothers [30]; however, the term Kalman smoother generally refers to the fixed-interval case in which the goal is to estimate the sequence of states over a finite time window using all observations in the same window.

In this paper, we assume that the parameters characterizing the second-order statistics of process and observation noise are unknown and propose an *optimal Bayesian Kalman smoother* (OBKS) framework to obtain smoothed estimates that are optimal relative to the posterior distribution of the unknown noise parameters. Referring to our method as an “optimal Bayesian” smoother is consistent with the terminology used in other works devoted to the design of optimal Bayesian filters when a prior distribution is assumed for the unknown parameters in the random process model [17, 18].

In a sense, this paper fills in the last block of a six-part Kalman filtering paradigm: (1) filter/predictor under known model, (2) smoother under known model, (3) adaptive filter/predictor under unknown model, (4) adaptive smoother under unknown model, (5) optimal filter/predictor relative to an uncertainty class of models, and (6) optimal smoother relative to an uncertainty class of models. This is not to say that all problems have been solved. There can be many adaptive approaches. There are also many ways in which there can be uncertainty in the state-space model, and optimality relative to that uncertainty can be defined via different cost functions. In the four uncertainty settings referred to here, the covariance matrices for the process and observation noise are assumed to be unknown (in a manner to be precisely defined in the sequel).

Similar to the IBR-KF and OBKF, the proposed smoother is rooted in an innovation process. Several ordinary Kalman smoothers have employed innovation processes: for continuous-time systems [31], the fixed-interval Kalman smoother for linear discrete-time systems when only the covariance information is available [32], and when observations might be randomly missing [33]. In this paper, we use the Bayesian innovation process and the Bayesian orthogonality principle proposed in [16] to derive the OBKS forward-backward equations. The main advantage of the proposed smoothing framework is that it possesses the same forward-backward structure as that of the ordinary Kalman smoother with the ordinary noise statistics replaced by their effective counterparts. The effective statistics incorporate the uncertainty of the parameters characterizing the observation and process noise second-order statistics in such a way that designing an OBKS relative to an uncertainty class can be reduced to designing an ordinary Kalman smoother relative to the effective statistics. Specifically, we introduce the *effective Kalman smoothing gain* for the backward step of the OBKS. The proposed smoothing framework requires

two forward steps. In the first step, the *posterior effective noise statistics* are computed. Then, the optimal Bayesian Kalman filter is designed relative to the obtained posterior effective noise statistics and is run in the forward direction over the window of observations. Finally, in the backward step, the Bayesian smoothed estimates are obtained.

This paper is organized as follows. In Section 2, we provide the theoretical foundation and derive the recursive equations for the proposed optimal Bayesian Kalman smoother. Section 3 is devoted to the experimental evaluation of the proposed OBKS method using two examples: target tracking and gene regulatory network inference. Finally, concluding remarks are given in Section 4.

Here, we summarize the notations employed throughout the paper. We use uppercase and lowercase boldface letters to denote matrices and vectors, respectively. \mathbf{M}^T , $|\mathbf{M}|$, and $\text{Tr}\{\mathbf{M}\}$ represent the transpose, determinant, and trace (sum of diagonal elements) operators for matrix \mathbf{M} , respectively. Also, $\text{diag}[\cdot]$ represents the diagonal elements of a diagonal matrix. The value of a time-dependent matrix at time k is denoted by \mathbf{M}_k . Let (P, Ω, \mathcal{E}) be a probability space, then $E[\cdot]$ denotes the expectation relative to the probability measure P . In a real-valued random vector $\mathbf{x} = [x(1), \dots, x(k)]$, each component is a real random variable $x(i) : \Omega \rightarrow \mathcal{R}$, $1 \leq i \leq k$. We use $E[\mathbf{x}]$ and $\text{cov}[\mathbf{x}] = E[\mathbf{x}\mathbf{x}^T]$ to denote the mean vector and the covariance matrix, respectively. Finally, $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ denotes a multivariate Gaussian function relative to random vector \mathbf{x} with the mean vector μ and the covariance matrix Σ .

2 Optimal Bayesian Kalman smoother

2.1 Problem formulation and theoretical background

In this paper, we consider the following parameterized state-space model:

$$\mathbf{x}_{k+1}^{\theta_1} = \Phi_k \mathbf{x}_k^{\theta_1} + \Gamma_k \mathbf{u}_k^{\theta_1} \tag{1}$$

$$\mathbf{y}_k^\theta = \mathbf{H}_k \mathbf{x}_k^{\theta_1} + \mathbf{v}_k^{\theta_2}, \tag{2}$$

where $\mathbf{x}_k^{\theta_1}$ and \mathbf{y}_k^θ are vectors of size $n \times 1$ and $m \times 1$, called the state vector and observation vector, respectively. Φ_k , \mathbf{H}_k , and Γ_k are matrices of size $n \times n$, $m \times n$, and $n \times p$ called the state transition matrix, observation transition matrix, and the process noise transition matrix, respectively. We let $\mathbf{z}_k^{\theta_1} = \mathbf{H}_k \mathbf{x}_k^{\theta_1}$. $\mathbf{u}_k^{\theta_1}$ and $\mathbf{v}_k^{\theta_2}$ are $p \times 1$ and $m \times 1$ vectors representing the process noise and observation noise, respectively, being zero-mean discrete white-noise processes. The unknown noise covariance matrices of the process and observation noise are given by

$$E \left[\mathbf{u}_k^{\theta_1} \left(\mathbf{u}_l^{\theta_1} \right)^T \right] = \mathbf{Q}^{\theta_1} \delta_{kl}, \tag{3}$$

$$E \left[\mathbf{v}_k^{\theta_2} \left(\mathbf{v}_l^{\theta_2} \right)^T \right] = \mathbf{R}^{\theta_2} \delta_{kl}, \tag{4}$$

where δ_{kl} is Dirac delta function, i.e., $\delta_{kl} = 1$ for $k = l$ and $\delta_{kl} = 0$ for $k \neq l$, and θ_1 and θ_2 are two unknown parameters such that $\boldsymbol{\theta} = [\theta_1, \theta_2] \in \Theta$, Θ being the collection of all possible realizations for $\boldsymbol{\theta}$, governed by a prior distribution $\pi(\boldsymbol{\theta})$. We assume that θ_1 and θ_2 are independent. Note that while the observation vector \mathbf{y}_k^θ depends on both θ_1 and θ_2 , the state vector $\mathbf{x}_k^{\theta_1}$ depends only on θ_1 .

Considering a state-space model according to (1) and (2) and an observation window $\mathcal{Y}_L^\theta = \{\mathbf{y}_0^\theta, \mathbf{y}_1^\theta, \dots, \mathbf{y}_L^\theta\}$ of size L , we desire an *optimal Bayesian Kalman smoother* (OBKS) that is a fixed-interval smoother involving finding the estimates of states $\mathbf{x}_0^{\theta_1}, \mathbf{x}_1^{\theta_1}, \dots, \mathbf{x}_L^{\theta_1}$ in the same window. In this context, the Bayesian smoothed estimate $\hat{\mathbf{x}}_{k|L}^\theta$ of $\mathbf{x}_k^{\theta_1}$, which is the output of the OBKS at time k , has the following form

$$\hat{\mathbf{x}}_{k|L}^\theta = \sum_{l=0}^L \mathbf{G}_{k,l}^\ominus \mathbf{y}_l^\theta, \tag{5}$$

such that the average MSE relative to the posterior distribution $\pi(\boldsymbol{\theta} | \mathcal{Y}_L^\theta)$ is minimized:

$$\mathbf{G}_{k,l}^\ominus = \arg \min_{\mathbf{G}_{k,l} \in \mathcal{G}} E_{\theta^*} \left[E \left[\left(\mathbf{x}_k^{\theta_1} - \sum_{l=0}^L \mathbf{G}_{k,l} \mathbf{y}_l^\theta \right)^T \times \left(\mathbf{x}_k^{\theta_1} - \sum_{l=0}^L \mathbf{G}_{k,l} \mathbf{y}_l^\theta \right) \right] \right], \tag{6}$$

where \mathcal{G} is the vector space of all $n \times m$ matrix-valued functions $\mathbf{G}_{k,l} : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}^{n \times m}$, and $E_{\theta^*}[\cdot]$ denotes the expectation relative to $\pi(\boldsymbol{\theta} | \mathcal{Y}_L^\theta)$, i.e., $E_{\theta^*}[\cdot] = \int_{\Theta} (\cdot) \pi(\boldsymbol{\theta} | \mathcal{Y}_L^\theta) d\boldsymbol{\theta}$. Note that we use $E_{\theta}[\cdot]$ to denote the expectation relative to the prior distribution $\pi(\boldsymbol{\theta})$. Furthermore, $E[\boldsymbol{\theta}]$ and $E[\boldsymbol{\theta} | \mathcal{Y}_L^\theta]$ represent the expectation of parameter $\boldsymbol{\theta}$ relative to $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta} | \mathcal{Y}_L^\theta)$, respectively. It is worth mentioning that the optimal Bayesian Kalman predictor and the optimal Bayesian Kalman filter proposed in [18] correspond to $L = k - 1$ and $L = k$ in (5), respectively. Also, if instead of $E_{\theta^*}[\cdot]$, $E_{\theta}[\cdot]$ is used in (6), the estimators corresponding to $L = k - 1$ and $L = k$ in (5) are called the intrinsically Bayesian robust Kalman predictor and filter, respectively [16].

Before developing the OBKS equations, we state a theorem and a lemma required for deriving equations whose proofs can be found in [16].

The next theorem is a restatement of the classical orthogonality principle relative to the inner product defined by $E_{\theta^*}[E[\cdot]]$ applied to $\mathbf{x}_k^{\theta_1}$, \mathbf{y}_l^θ , and $\hat{\mathbf{x}}_{k|L}^\theta$, keeping in mind that $\mathbf{x}_k^{\theta_1}$ depends only on θ_1 , whereas $\hat{\mathbf{x}}_{k|L}^\theta$ depends on $\boldsymbol{\theta} = [\theta_1, \theta_2]$. As originally stated in [16], the ‘‘Bayesian orthogonality principle’’ involved the inner product defined by $E_{\theta}[E[\cdot]]$.

Theorem 1 (Bayesian Orthogonality Principle) *The Bayesian smoothed estimate obtained in (5) satisfies (6) (having minimum average MSE relative to the posterior distribution) if and only if*

$$E_{\theta^*} \left[E \left[\left(\mathbf{x}_k^{\theta_1} - \widehat{\mathbf{x}}_{k|L}^{\theta} \right) \left(\mathbf{y}_l^{\theta} \right)^T \right] \right] = \mathbf{0}_{n \times m}, \quad (7)$$

for all $l \leq L$, where $\mathbf{0}_{n \times m}$ is the zero matrix of size $n \times m$.

If $\widehat{\mathbf{x}}_{k|k-1}^{\theta}$ is the output of the optimal Bayesian Kalman predictor at time k , then the *Bayesian innovation process* is defined as [16]

$$\widetilde{\mathbf{z}}_k^{\theta} = \mathbf{y}_k^{\theta} - \mathbf{H}_k \widehat{\mathbf{x}}_{k|k-1}^{\theta}. \quad (8)$$

It can be shown that [16]

$$E_{\theta^*} \left[E \left[\widetilde{\mathbf{z}}_k^{\theta} \left(\widetilde{\mathbf{z}}_l^{\theta} \right)^T \right] \right] = E_{\theta^*} \left[\mathbf{H}_k \mathbf{P}_{k|k-1}^{\mathbf{x}, \theta} \mathbf{H}_k^T + \mathbf{R}^{\theta_2} \right] \delta_{kl}, \quad (9)$$

where

$$\mathbf{P}_{k|k-1}^{\mathbf{x}, \theta} = E \left[\left(\mathbf{x}_k^{\theta_1} - \widehat{\mathbf{x}}_{k|k-1}^{\theta} \right) \left(\mathbf{x}_k^{\theta_1} - \widehat{\mathbf{x}}_{k|k-1}^{\theta} \right)^T \right], \quad (10)$$

is the Bayesian prediction error covariance matrix relative to θ . Note that if $\mathbf{z}_k^{\theta_1} = \mathbf{H}_k \mathbf{x}_k^{\theta_1}$ and $\widehat{\mathbf{z}}_k^{\theta} = \mathbf{H}_k \widehat{\mathbf{x}}_k^{\theta}$, then

$$\begin{aligned} \mathbf{P}_{k|k-1}^{\mathbf{z}, \theta} &= E \left[\left(\mathbf{z}_k^{\theta_1} - \widehat{\mathbf{z}}_{k|k-1}^{\theta} \right) \left(\mathbf{z}_k^{\theta_1} - \widehat{\mathbf{z}}_{k|k-1}^{\theta} \right)^T \right] \\ &= \mathbf{H}_k \mathbf{P}_{k|k-1}^{\mathbf{x}, \theta} \mathbf{H}_k^T. \end{aligned} \quad (11)$$

The following lemma, which can be proved similar to the proof given in [16], helps us find the Bayesian smoothed estimates using the Bayesian innovation process.

Lemma 1 (Bayesian Information Equivalence) *The Bayesian smoothed estimate $\widehat{\mathbf{x}}_{k|L}^{\theta}$ for \mathbf{x}_k^{θ} based upon observations \mathbf{y}_l^{θ} , $0 \leq l \leq L$, can be found by computing the Bayesian smoothed estimate based upon the Bayesian innovation process $\widetilde{\mathbf{z}}_l^{\theta}$, $0 \leq l \leq L$.*

2.2 Update equation for Bayesian smoothed estimate

We now proceed to develop the recursive structure of the OBKS based on the theoretical foundation laid out in the previous subsection.

Using Lemma 1, we can have the following form for the Bayesian smoothed estimate $\widehat{\mathbf{x}}_{k|L}^{\theta}$ defined in (5):

$$\widehat{\mathbf{x}}_{k|L}^{\theta} = \sum_{l=0}^L \mathbf{G}_{k,l}^{\ominus} \widetilde{\mathbf{z}}_l^{\theta}, \quad (12)$$

where $\widehat{\mathbf{x}}_{k|L}^{\theta}$ obtained in (12) should satisfy the Bayesian orthogonality principle (relative to $\widetilde{\mathbf{z}}_l^{\theta}$), for $l \leq L$,

$$E_{\theta^*} \left[E \left[\left(\mathbf{x}_k^{\theta_1} - \widehat{\mathbf{x}}_{k|L}^{\theta} \right) \left(\widetilde{\mathbf{z}}_l^{\theta} \right)^T \right] \right] = \mathbf{0}_{n \times m}. \quad (13)$$

After some mathematical manipulations and also using (9), one can verify that

$$\mathbf{G}_{k,l}^{\ominus} = E_{\theta^*} \left[E \left[\mathbf{x}_k^{\theta_1} \left(\widetilde{\mathbf{z}}_l^{\theta} \right)^T \right] \right] E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{\mathbf{z}, \theta} + \mathbf{R}^{\theta_2} \right]. \quad (14)$$

Plugging (14) in (12) yields

$$\begin{aligned} \widehat{\mathbf{x}}_{k|L}^{\theta} &= \sum_{l=0}^L E_{\theta^*} \left[E \left[\mathbf{x}_k^{\theta_1} \left(\widetilde{\mathbf{z}}_l^{\theta} \right)^T \right] \right] E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{\mathbf{z}, \theta} + \mathbf{R}^{\theta_2} \right] \widetilde{\mathbf{z}}_l^{\theta} \\ &= \widehat{\mathbf{x}}_{k|k}^{\theta} + \sum_{l=k+1}^L E_{\theta^*} \left[E \left[\mathbf{x}_k^{\theta_1} \left(\widetilde{\mathbf{z}}_l^{\theta} \right)^T \right] \right] E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{\mathbf{z}, \theta} + \mathbf{R}^{\theta_2} \right] \widetilde{\mathbf{z}}_l^{\theta}, \end{aligned} \quad (15)$$

where $\widehat{\mathbf{x}}_{k|k}^{\theta}$ is the output of the OBKF, developed in [18], at time k . We can further simplify $E_{\theta^*} \left[E \left[\mathbf{x}_k^{\theta_1} \left(\widetilde{\mathbf{z}}_l^{\theta} \right)^T \right] \right]$, for $k+1 \leq l \leq L$, as

$$\begin{aligned} E_{\theta^*} \left[E \left[\mathbf{x}_k^{\theta_1} \left(\widetilde{\mathbf{z}}_l^{\theta} \right)^T \right] \right] &= E_{\theta^*} \left[E \left[\left(\widehat{\mathbf{x}}_k^{\theta} + \widehat{\mathbf{x}}_{k|k-1}^{\theta} \right) \left(\mathbf{H}_l \mathbf{x}_l^{\theta_1} + \mathbf{v}_l^{\theta_2} - \mathbf{H}_l \widehat{\mathbf{x}}_{l|l-1}^{\theta} \right)^T \right] \right] \\ &= E_{\theta^*} \left[E \left[\left(\widehat{\mathbf{x}}_k^{\theta} + \widehat{\mathbf{x}}_{k|k-1}^{\theta} \right) \left(\mathbf{H}_l \widehat{\mathbf{x}}_l^{\theta} + \mathbf{v}_l^{\theta_2} \right)^T \right] \right] \\ &= E_{\theta^*} \left[\mathbf{W}_{k,l}^{\theta} \right] \mathbf{H}_l^T, \end{aligned} \quad (16)$$

where $\widehat{\mathbf{x}}_k^{\theta} = \mathbf{x}_k^{\theta_1} - \widehat{\mathbf{x}}_{k|k-1}^{\theta}$ is the Bayesian prediction error relative to θ at time k and its auto-correlation $E \left[\widehat{\mathbf{x}}_k^{\theta} \left(\widehat{\mathbf{x}}_l^{\theta} \right)^T \right]$ is denoted by $\mathbf{W}_{k,l}^{\theta}$. Note that the third equality in (16) results from the fact that $E_{\theta^*} \left[E \left[\widehat{\mathbf{x}}_{k|k-1}^{\theta} \left(\widehat{\mathbf{x}}_l^{\theta} \right)^T \right] \right] = \mathbf{0}_{n \times n}$ due to the Bayesian orthogonality principle and $\mathbf{v}_l^{\theta_2}$ is independent from $\widehat{\mathbf{x}}_k^{\theta}$ and $\widehat{\mathbf{x}}_{k|k-1}^{\theta}$ for $k < l$. Therefore, substituting (16), (15) becomes

$$\widehat{\mathbf{x}}_{k|L}^{\theta} = \widehat{\mathbf{x}}_{k|k}^{\theta} + \sum_{l=k+1}^L E_{\theta^*} \left[\mathbf{W}_{k,l}^{\theta} \right] \mathbf{H}_l^T E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{\mathbf{z}, \theta} + \mathbf{R}^{\theta_2} \right] \widetilde{\mathbf{z}}_l^{\theta}. \quad (17)$$

Keeping in mind that we want to find a backward recursive formulation for $\tilde{\mathbf{x}}_{k|L}^\theta$, writing (15) for $k+1$ we have

$$\begin{aligned}\tilde{\mathbf{x}}_{k+1|L}^\theta &= \sum_{l=0}^k E_{\theta^*} \left[E \left[\mathbf{x}_{k+1}^{\theta_1} (\tilde{\mathbf{z}}_l^\theta)^T \right] \right] E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{z,\theta} + \mathbf{R}^{\theta_2} \right] \tilde{\mathbf{z}}_l^\theta \\ &\quad + \sum_{l=k+1}^L E_{\theta^*} \left[E \left[\mathbf{x}_{k+1}^{\theta_1} (\tilde{\mathbf{z}}_l^\theta)^T \right] \right] E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{z,\theta} + \mathbf{R}^{\theta_2} \right] \tilde{\mathbf{z}}_l^\theta \\ &= \tilde{\mathbf{x}}_{k+1|k}^\theta + \sum_{l=k+1}^L E_{\theta^*} \left[E \left[\mathbf{x}_{k+1}^{\theta_1} (\tilde{\mathbf{z}}_l^\theta)^T \right] \right] E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{z,\theta} + \mathbf{R}^{\theta_2} \right] \tilde{\mathbf{z}}_l^\theta \\ &= \tilde{\mathbf{x}}_{k+1|k}^\theta + \sum_{l=k+1}^L E_{\theta^*} \left[\mathbf{W}_{k+1,l}^\theta \right] \mathbf{H}_l^T E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{z,\theta} + \mathbf{R}^{\theta_2} \right] \tilde{\mathbf{z}}_l^\theta.\end{aligned}\quad (18)$$

Considering (17) and (18), one can conclude that obtaining an update equation for $\tilde{\mathbf{x}}_{k|L}^\theta$ requires a recursive relationship between $E_{\theta^*} \left[\mathbf{W}_{k,l}^\theta \right]$ and $E_{\theta^*} \left[\mathbf{W}_{k+1,l}^\theta \right]$ and since $\mathbf{W}_{k,l}^\theta$ involves the Bayesian prediction error $\tilde{\mathbf{x}}_k^\theta$, the first step is to find the update equation for $\tilde{\mathbf{x}}_k^\theta$. As shown in [16],

$$\tilde{\mathbf{x}}_{k+1}^\theta = \bar{\Phi}_k^\ominus \tilde{\mathbf{x}}_k^\theta + \Gamma_k \mathbf{u}_k^{\theta_1} - \Phi_k \mathbf{K}_k^\ominus \mathbf{v}_k^{\theta_2}, \quad (19)$$

in which

$$\bar{\Phi}_k^\ominus = \Phi_k (\mathbf{I} - \mathbf{K}_k^\ominus \mathbf{H}_k), \quad (20)$$

and

$$\mathbf{K}_k^\ominus = E_{\theta^*} \left[\mathbf{P}_{k|k-1}^{x,\theta} \right] \mathbf{H}_k^T E_{\theta^*}^{-1} \left[\mathbf{P}_{k|k-1}^{z,\theta} + \mathbf{R}^{\theta_2} \right], \quad (21)$$

is called the *effective Kalman gain matrix* [18]. Also, we call $E_{\theta^*} \left[\mathbf{Q}^{\theta_1} \right]$ and $E_{\theta^*} \left[\mathbf{R}^{\theta_2} \right]$ the *posterior effective process noise statistics* and the *posterior effective observation noise statistics*, respectively. As has been shown in [18], $E_{\theta^*} \left[\mathbf{Q}^{\theta_1} \right]$ is required for updating $E_{\theta^*} \left[\mathbf{P}_{k|k-1}^{z,\theta} \right]$. Using (19), we find the relation between $\tilde{\mathbf{x}}_l^\theta$ and $\tilde{\mathbf{x}}_k^\theta$ as follows:

$$\begin{aligned}\tilde{\mathbf{x}}_l^\theta &= \bar{\Phi}_{l-1}^\ominus \tilde{\mathbf{x}}_{l-1}^\theta + \Gamma_{l-1} \mathbf{u}_{l-1}^{\theta_1} - \Phi_{l-1} \mathbf{K}_{l-1}^\ominus \mathbf{v}_{l-1}^{\theta_2} \\ &= \bar{\Phi}_{l-1}^\ominus \bar{\Phi}_{l-2}^\ominus \tilde{\mathbf{x}}_{l-2}^\theta + \bar{\Phi}_{l-1}^\ominus \left(\Gamma_{l-2} \mathbf{u}_{l-2}^{\theta_1} - \Phi_{l-2} \mathbf{K}_{l-2}^\ominus \mathbf{v}_{l-2}^{\theta_2} \right) \\ &\quad + \left(\Gamma_{l-1} \mathbf{u}_{l-1}^{\theta_1} - \Phi_{l-1} \mathbf{K}_{l-1}^\ominus \mathbf{v}_{l-1}^{\theta_2} \right) \\ &\vdots \\ &= \bar{\Phi}_{l-1}^\ominus \bar{\Phi}_{l-2}^\ominus \dots \bar{\Phi}_k^\ominus \tilde{\mathbf{x}}_k^\theta + \sum_{l=k}^{l-1} \bar{\Phi}_{l-1}^\ominus \bar{\Phi}_{l-2}^\ominus \dots \bar{\Phi}_{l+1}^\ominus \left(\Gamma_l \mathbf{u}_l^{\theta_1} - \Phi_l \mathbf{K}_l^\ominus \mathbf{v}_l^{\theta_2} \right).\end{aligned}\quad (22)$$

Using (22), we have

$$\begin{aligned}E_{\theta^*} \left[\mathbf{W}_{k,l}^\theta \right] &= E_{\theta^*} \left[E \left[\tilde{\mathbf{x}}_k^\theta (\tilde{\mathbf{z}}_l^\theta)^T \right] \right] \\ &= E_{\theta^*} \left[E \left[\tilde{\mathbf{x}}_k^\theta (\tilde{\mathbf{x}}_k^\theta)^T \right] \right] \left(\bar{\Phi}_k^\ominus \right)^T \left(\bar{\Phi}_{k+1}^\ominus \right)^T \dots \left(\bar{\Phi}_{l-1}^\ominus \right)^T \\ &= E_{\theta^*} \left[\mathbf{P}_{k|k-1}^{x,\theta} \right] (\mathbf{I} - \mathbf{K}_k^\ominus \mathbf{H}_k)^T \Phi_k^T \dots (\mathbf{I} - \mathbf{K}_{l-1}^\ominus \mathbf{H}_{l-1})^T \Phi_{l-1}^T,\end{aligned}\quad (23)$$

where the second equality is obtained because, for $l' \geq k$, future process noise $\mathbf{u}_{l'}^{\theta_1}$ and observation noise $\mathbf{v}_{l'}^{\theta_2}$ are independent from $\tilde{\mathbf{x}}_k^\theta$. Now plugging

$$E_{\theta^*} \left[\mathbf{P}_{k|k}^{x,\theta} \right] = (\mathbf{I} - \mathbf{K}_k^\ominus \mathbf{H}_k) E_{\theta^*} \left[\mathbf{P}_{k|k-1}^{x,\theta} \right], \quad (24)$$

derived in [16], in (23) yields the following recursive equation for $E_{\theta^*} \left[\mathbf{W}_{k,l}^\theta \right]$:

$$\begin{aligned}E_{\theta^*} \left[\mathbf{W}_{k,l}^\theta \right] &= E_{\theta^*} \left[\mathbf{P}_{k|k}^{x,\theta} \right] \Phi_k^T (\mathbf{I} - \mathbf{K}_{k+1}^\ominus \mathbf{H}_{k+1})^T \Phi_{k+1}^T \dots (\mathbf{I} - \mathbf{K}_{l-1}^\ominus \mathbf{H}_{l-1})^T \Phi_{l-1}^T \\ &= E_{\theta^*} \left[\mathbf{P}_{k|k}^{x,\theta} \right] \Phi_k^T E_{\theta^*}^{-1} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right] \left(E_{\theta^*} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right] (\mathbf{I} - \mathbf{K}_{k+1}^\ominus \mathbf{H}_{k+1})^T \Phi_{k+1}^T \right. \\ &\quad \left. \dots (\mathbf{I} - \mathbf{K}_{l-1}^\ominus \mathbf{H}_{l-1})^T \Phi_{l-1}^T \right) \\ &= E_{\theta^*} \left[\mathbf{P}_{k|k}^{x,\theta} \right] \Phi_k^T E_{\theta^*}^{-1} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right] E_{\theta^*} \left[\mathbf{W}_{k+1,l}^\theta \right],\end{aligned}\quad (25)$$

where the second equality is obtained by multiplying $E_{\theta^*}^{-1} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right] E_{\theta^*} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right]$ in the first equality and (23) written for $E_{\theta^*} \left[\mathbf{W}_{k+1,l}^\theta \right]$ is used to obtain the last equality. Substituting (25) in (17) yields the update equation for $\tilde{\mathbf{x}}_{k|L}^\theta$ as

$$\begin{aligned}\tilde{\mathbf{x}}_{k|L}^\theta &= \tilde{\mathbf{x}}_{k|k}^\theta + \sum_{l=k+1}^L E_{\theta^*} \left[\mathbf{P}_{k|k}^{x,\theta} \right] \Phi_k^T E_{\theta^*}^{-1} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right] \\ &\quad \times E_{\theta^*} \left[\mathbf{W}_{k+1,l}^\theta \right] \mathbf{H}_l^T E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{z,\theta} + \mathbf{R}^{\theta_2} \right] \tilde{\mathbf{z}}_l^\theta \\ &= \tilde{\mathbf{x}}_{k|k}^\theta + E_{\theta^*} \left[\mathbf{P}_{k|k}^{x,\theta} \right] \Phi_k^T E_{\theta^*}^{-1} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right] \sum_{l=k+1}^L \left(E_{\theta^*} \left[\mathbf{W}_{k+1,l}^\theta \right] \right. \\ &\quad \left. \times \mathbf{H}_l^T E_{\theta^*}^{-1} \left[\mathbf{P}_{l|l-1}^{z,\theta} + \mathbf{R}^{\theta_2} \right] \tilde{\mathbf{z}}_l^\theta \right) \\ &= \tilde{\mathbf{x}}_{k|k}^\theta + \mathbf{A}_k^\ominus \left(\tilde{\mathbf{x}}_{k+1|L}^\theta - \tilde{\mathbf{x}}_{k+1|k}^\theta \right),\end{aligned}\quad (26)$$

where the third equality is obtained due to (18) and

$$\mathbf{A}_k^\ominus = E_{\theta^*} \left[\mathbf{P}_{k|k}^{x,\theta} \right] \Phi_k^T E_{\theta^*}^{-1} \left[\mathbf{P}_{k+1|k}^{x,\theta} \right], \quad (27)$$

is called the *effective Kalman smoothing gain*.

2.3 Update equation for the Bayesian smoothed error covariance matrix

Letting $\tilde{\mathbf{x}}_{k|L}^{\theta_1} = \mathbf{x}_k^{\theta_1} - \tilde{\mathbf{x}}_{k|L}^\theta$ be the Bayesian smoothed error relative to θ at time k , we now aim to find a recursive formulation for the average Bayesian smoothing error covariance matrix

$$E_{\theta^*} \left[\mathbf{P}_{k|L}^{x,\theta} \right] = E_{\theta^*} \left[E \left[\tilde{\mathbf{x}}_{k|L}^{\theta_1} (\tilde{\mathbf{x}}_{k|L}^{\theta_1})^T \right] \right]. \quad (28)$$

To do so, we first obtain the following relation for $\tilde{\mathbf{x}}_{k|L}^\theta$:

$$\begin{aligned}
\tilde{\mathbf{x}}_{k|L}^\theta &= \mathbf{x}_k^{\theta_1} - \tilde{\mathbf{x}}_{k|L}^\theta \\
&= \mathbf{x}_k^{\theta_1} - \tilde{\mathbf{x}}_{k|k}^\theta - \mathbf{A}_k^\ominus \left(\tilde{\mathbf{x}}_{k+1|L}^\theta - \tilde{\mathbf{x}}_{k+1|k}^\theta \right) \\
&= \tilde{\mathbf{x}}_{k|k}^\theta - \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|L}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|k}^\theta \\
&= \tilde{\mathbf{x}}_{k|k}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|L}^\theta - \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|k}^\theta.
\end{aligned} \quad (29)$$

Hence,

$$\tilde{\mathbf{x}}_{k|L}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|k}^\theta = \tilde{\mathbf{x}}_{k|k}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|L}^\theta. \quad (30)$$

Taking the covariance matrix of both sides of (30) relative to $E_{\theta^*}[E[\cdot]]$ yields

$$\begin{aligned}
E_{\theta^*} \left[E \left[\left(\tilde{\mathbf{x}}_{k|L}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|k}^\theta \right) \left(\tilde{\mathbf{x}}_{k|L}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|k}^\theta \right)^T \right] \right] \\
= E_{\theta^*} \left[E \left[\left(\tilde{\mathbf{x}}_{k|k}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|L}^\theta \right) \left(\tilde{\mathbf{x}}_{k|k}^\theta + \mathbf{A}_k^\ominus \tilde{\mathbf{x}}_{k+1|L}^\theta \right)^T \right] \right].
\end{aligned} \quad (31)$$

Due to the Bayesian orthogonality principle, for the left-hand side of (31),

$$E_{\theta^*} \left[E \left[\tilde{\mathbf{x}}_{k|L}^\theta \left(\tilde{\mathbf{x}}_{k+1|k}^\theta \right)^T \right] \right] = \mathbf{0}_{n \times n}. \quad (32)$$

Similarly, regarding the right-hand side of (31),

$$E_{\theta^*} \left[E \left[\tilde{\mathbf{x}}_{k|k}^\theta \left(\tilde{\mathbf{x}}_{k+1|L}^\theta \right)^T \right] \right] = \mathbf{0}_{n \times n}. \quad (33)$$

Thus, (31) can be simplified to

$$\begin{aligned}
E_{\theta^*} \left[\mathbf{P}_{k|L}^{\mathbf{x}, \theta} \right] + \mathbf{A}_k^\ominus E_{\theta^*} \left[\mathbf{P}_{k+1|k}^{\mathbf{x}, \theta} \right] \left(\mathbf{A}_k^\ominus \right)^T \\
= E_{\theta^*} \left[\mathbf{P}_{k|k}^{\mathbf{x}, \theta} \right] + \mathbf{A}_k^\ominus E_{\theta^*} \left[\mathbf{P}_{k+1|L}^{\mathbf{x}, \theta} \right] \left(\mathbf{A}_k^\ominus \right)^T.
\end{aligned} \quad (34)$$

Finally, (34) can be rearranged to obtain an update equation for $E_{\theta^*} \left[\mathbf{P}_{k|L}^{\mathbf{x}, \theta} \right]$

$$\begin{aligned}
E_{\theta^*} \left[\mathbf{P}_{k|L}^{\mathbf{x}, \theta} \right] = E_{\theta^*} \left[\mathbf{P}_{k|k}^{\mathbf{x}, \theta} \right] + \mathbf{A}_k^\ominus \left(E_{\theta^*} \left[\mathbf{P}_{k+1|L}^{\mathbf{x}, \theta} \right] \right. \\
\left. - E_{\theta^*} \left[\mathbf{P}_{k+1|k}^{\mathbf{x}, \theta} \right] \right) \left(\mathbf{A}_k^\ominus \right)^T.
\end{aligned} \quad (35)$$

Finding the average Bayesian smoothing error covariance matrix in (35) completes all equations needed for implementing the OBKS framework.

The forward step of the OBKS involves running the OBKF and in the backward step the Bayesian smoothed estimates are obtained. We should point out that in practice for the OBKF, the posterior effective noise statistics are updated sequentially for each k because filtering is an online estimation scheme. However, here since we use OBKF as the forward step of the OBKS, which is an offline estimation, we use the posterior effective noise statistics $E[\theta|\mathcal{Y}_L^\theta]$ relative to the whole observation window for the OBKF-based estimation from the beginning. In other words, the OBKF used in the forward step, is in fact

the IBR-KF designed relative to the posterior distribution $\pi(\theta|\mathcal{Y}_L^\theta)$.

To better understand the similarity between the recursive structures of the proposed OBKS and the ordinary Kalman smoother, Table 1 compares the recursive equations required for these two smoothers. As this table suggests, the recursive structure of the proposed OBKS framework is similar to that of the classical Kalman smoother except that it employs effective characteristics, namely, the effective Kalman gain matrix \mathbf{K}_k^\ominus , effective Kalman smoothing gain matrix \mathbf{A}_k^\ominus , and the posterior effective noise statistics $E_{\theta^*}[\mathbf{Q}^{\theta_1}]$ and $E_{\theta^*}[\mathbf{R}^{\theta_2}]$.

If the state vector \mathbf{x}_0 is characterized by $E[\mathbf{x}_0]$ and $\text{cov}[\mathbf{x}_0]$, then the forward step of the OBKS is initialized as $E_{\theta^*}[\mathbf{P}_{0|0}^{\mathbf{x}, \theta}] = \text{cov}[\mathbf{x}_0]$, $E_{\theta^*}[\mathbf{P}_{1|0}^{\mathbf{x}, \theta}] = \Phi_0 \text{cov}[\mathbf{x}_0] \Phi_0^T + \Gamma_0 E_{\theta^*}[\mathbf{Q}^{\theta_1}] \Gamma_0^T$, $\tilde{\mathbf{x}}_{0|0}^\theta = E[\mathbf{x}_0]$, and $\tilde{\mathbf{x}}_{1|0}^\theta = \Phi_0 E[\mathbf{x}_0]$.

2.4 Computing posterior effective noise statistics

The forward step in the OBKS requires the posterior effective noise statistics $E_{\theta^*}[\mathbf{Q}^{\theta_1}]$ and $E_{\theta^*}[\mathbf{R}^{\theta_2}]$. These expectations should be computed relative to the posterior distribution $\pi(\theta|\mathcal{Y}_L^\theta)$. Since the closed-form solution

Table 1 Comparison of the recursive equations for the classical and the proposed optimal Bayesian Kalman smoothers

	Forward step ($k = 1, \dots, L$)
Classical Kalman smoother	$\tilde{\mathbf{z}}_k = \mathbf{y}_k - \mathbf{H}_k \tilde{\mathbf{x}}_{k k-1}$ $\mathbf{K}_k = \mathbf{P}_{k k-1}^{\mathbf{x}} \mathbf{H}_k^T \left(\mathbf{H}_k \mathbf{P}_{k k-1}^{\mathbf{x}} \mathbf{H}_k^T + \mathbf{R} \right)^{-1}$ $\hat{\mathbf{x}}_{k k} = \tilde{\mathbf{x}}_{k k-1} + \mathbf{K}_k \tilde{\mathbf{z}}_k$ $\hat{\mathbf{x}}_{k+1 k} = \Phi_k \hat{\mathbf{x}}_{k k-1} + \Phi_k \mathbf{K}_k \tilde{\mathbf{z}}_k$ $\mathbf{P}_{k k}^{\mathbf{x}} = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k k-1}^{\mathbf{x}}$ $\mathbf{P}_{k+1 k}^{\mathbf{x}} = \Phi_k (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k k-1}^{\mathbf{x}} \Phi_k^T + \Gamma_k \mathbf{Q} \Gamma_k^T$
Optimal Bayesian Kalman smoother	$\tilde{\mathbf{z}}_k^\theta = \mathbf{y}_k^\theta - \mathbf{H}_k \tilde{\mathbf{x}}_{k k-1}^\theta$ $\mathbf{K}_k^\ominus = E_{\theta^*} \left[\mathbf{P}_{k k-1}^{\mathbf{x}, \theta} \right] \mathbf{H}_k^T E_{\theta^*}^{-1} \left[\mathbf{H}_k \mathbf{P}_{k k-1}^{\mathbf{x}, \theta} \mathbf{H}_k^T + \mathbf{R}^{\theta_2} \right]$ $\hat{\mathbf{x}}_{k k}^\theta = \tilde{\mathbf{x}}_{k k-1}^\theta + \mathbf{K}_k^\ominus \tilde{\mathbf{z}}_k^\theta$ $\hat{\mathbf{x}}_{k+1 k}^\theta = \Phi_k \hat{\mathbf{x}}_{k k-1}^\theta + \Phi_k \mathbf{K}_k^\ominus \tilde{\mathbf{z}}_k^\theta$ $E_{\theta^*} \left[\mathbf{P}_{k k}^{\mathbf{x}, \theta} \right] = (\mathbf{I} - \mathbf{K}_k^\ominus \mathbf{H}_k) E_{\theta^*} \left[\mathbf{P}_{k k-1}^{\mathbf{x}, \theta} \right]$ $E_{\theta^*} \left[\mathbf{P}_{k+1 k}^{\mathbf{x}, \theta} \right] = \Phi_k (\mathbf{I} - \mathbf{K}_k^\ominus \mathbf{H}_k) E_{\theta^*} \left[\mathbf{P}_{k k-1}^{\mathbf{x}, \theta} \right] \Phi_k^T + \Gamma_k E_{\theta^*} \left[\mathbf{Q}^{\theta_1} \right] \Gamma_k^T$
	Backward step ($k = L-1, L-2, \dots, 0$)
Classical Kalman smoother	$\mathbf{A}_k = \mathbf{P}_{k k}^{\mathbf{x}} \Phi_k^T \left(\mathbf{P}_{k+1 k}^{\mathbf{x}} \right)^{-1}$ $\hat{\mathbf{x}}_{k L} = \hat{\mathbf{x}}_{k+1 L} + \mathbf{A}_k \left(\hat{\mathbf{x}}_{k+1 L} - \hat{\mathbf{x}}_{k+1 k} \right)$ $\mathbf{P}_{k L}^{\mathbf{x}} = \mathbf{P}_{k+1 L}^{\mathbf{x}} + \mathbf{A}_k \left(\mathbf{P}_{k+1 L}^{\mathbf{x}} - \mathbf{P}_{k+1 k}^{\mathbf{x}} \right) \mathbf{A}_k^T$
Optimal Bayesian Kalman smoother	$\mathbf{A}_k^\ominus = E_{\theta^*} \left[\mathbf{P}_{k k}^{\mathbf{x}, \theta} \right] \Phi_k^T E_{\theta^*}^{-1} \left[\mathbf{P}_{k+1 k}^{\mathbf{x}, \theta} \right]$ $\hat{\mathbf{x}}_{k L}^\theta = \hat{\mathbf{x}}_{k+1 L}^\theta + \mathbf{A}_k^\ominus \left(\hat{\mathbf{x}}_{k+1 L}^\theta - \hat{\mathbf{x}}_{k+1 k}^\theta \right)$ $E_{\theta^*} \left[\mathbf{P}_{k L}^{\mathbf{x}, \theta} \right] = E_{\theta^*} \left[\mathbf{P}_{k+1 L}^{\mathbf{x}, \theta} \right] + \mathbf{A}_k^\ominus \left(E_{\theta^*} \left[\mathbf{P}_{k+1 k}^{\mathbf{x}, \theta} \right] \right. \\ \left. - E_{\theta^*} \left[\mathbf{P}_{k+1 L}^{\mathbf{x}, \theta} \right] \right) \left(\mathbf{A}_k^\ominus \right)^T$

of the distribution is not available, these expectations can be approximated via a Metropolis Hastings Markov chain Monte Carlo (MCMC) approach as proposed in [18]. To implement the MCMC method, we need to compute the likelihood function $f(\mathcal{Y}_L^\theta | \theta)$. Here, we outline the main steps needed to compute the likelihood function and refer to [18] for more details. Taking into account the Markov assumptions in the state-space models, the likelihood function can be written as

$$f(\mathcal{Y}_L^\theta | \theta) = \int \dots \int \left(\prod_{i=0}^L f(y_i | \mathbf{x}_i, \theta) \times \prod_{i=1}^L f(\mathbf{x}_i | \mathbf{x}_{i-1}, \theta) f(\mathbf{x}_0) \right) d\mathbf{x}_0 \dots d\mathbf{x}_L. \quad (36)$$

Letting $\tilde{\mathbf{Q}}_k^{\theta_1} = \Gamma_k \mathbf{Q}^{\theta_1} \Gamma_k^T$, since

$$f(\mathbf{y}_i | \mathbf{x}_i, \theta) = \mathcal{N}(\mathbf{y}_i; \mathbf{H}_i \mathbf{x}_i, \mathbf{R}^{\theta_2}), \quad (37)$$

$$f(\mathbf{x}_i | \mathbf{x}_{i-1}, \theta) = \mathcal{N}(\mathbf{x}_i; \Phi_{i-1} \mathbf{x}_{i-1}, \tilde{\mathbf{Q}}_k^{\theta_1}), \quad (38)$$

$f(\mathcal{Y}_L^\theta | \theta)$ can be regarded as a factorization of a global function for which we can use a sum-product algorithm called the factor graph [34]. Utilizing factor graphs, it can be seen that the likelihood function can be obtained as [18]

$$f(\mathcal{Y}_L^\theta | \theta) = S_L \sqrt{\frac{|\Delta_L|}{|\Sigma_L|}} \mathcal{N}(\mathbf{y}_L; \mathbf{0}_{m \times 1}, \mathbf{R}^{\theta_2}) \times \exp\left(\frac{1}{2} \left(\mathbf{G}_L^T \Delta_L^{-1} \mathbf{G}_L - \mathbf{M}_L^T \Sigma_L^{-1} \mathbf{M}_L \right)\right), \quad (39)$$

where

$$\Delta_L^{-1} = \mathbf{H}_L^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{H}_L + \Sigma_L^{-1}, \quad (40)$$

$$\mathbf{G}_L = \Delta_L \left(\mathbf{H}_L^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_L + \Sigma_L^{-1} \mathbf{M}_L \right), \quad (41)$$

and Σ_L , \mathbf{M}_L , and S_L are computed recursively utilizing the following expressions from $k = 0$ to $k = L - 1$:

$$\Sigma_{k+1}^{-1} = \left(\tilde{\mathbf{Q}}_k^{\theta_1} \right)^{-1} - \left(\tilde{\mathbf{Q}}_k^{\theta_1} \right)^{-1} \Phi_k \Lambda_k \Phi_k^T \left(\tilde{\mathbf{Q}}_k^{\theta_1} \right)^{-1} \quad (42)$$

$$\mathbf{M}_{k+1} = \Sigma_{k+1} \left(\tilde{\mathbf{Q}}_k^{\theta_1} \right)^{-1} \Phi_k \Lambda_k \left(\mathbf{H}_k^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_k + \Sigma_k^{-1} \mathbf{M}_k \right) \quad (43)$$

$$S_{k+1} = S_k \sqrt{\frac{|\Lambda_k| |\Sigma_{k+1}|}{|\tilde{\mathbf{Q}}_k^{\theta_1}| |\Sigma_k|}} \mathcal{N}(\mathbf{y}_k; \mathbf{0}_{m \times 1}, \mathbf{R}^{\theta_2}) \times \exp\left(\frac{\mathbf{M}_{k+1}^T \Sigma_{k+1}^{-1} \mathbf{M}_{k+1} + \mathbf{W}_k^T \Lambda_k \mathbf{W}_k - \mathbf{M}_k^T \Sigma_k^{-1} \mathbf{M}_k}{2}\right), \quad (44)$$

where Λ_k and \mathbf{W}_k are obtained as

$$\Lambda_k = \left(\Phi_k^T \left(\tilde{\mathbf{Q}}_k^{\theta_1} \right)^{-1} \Phi_k + \mathbf{H}_k (\mathbf{R}^{\theta_2})^{-1} \mathbf{H}_k + \Sigma_k^{-1} \right)^{-1}, \quad (45)$$

$$\mathbf{W}_k = \mathbf{H}_k^T (\mathbf{R}^{\theta_2})^{-1} \mathbf{y}_k + \Sigma_k^{-1} \mathbf{M}_k. \quad (46)$$

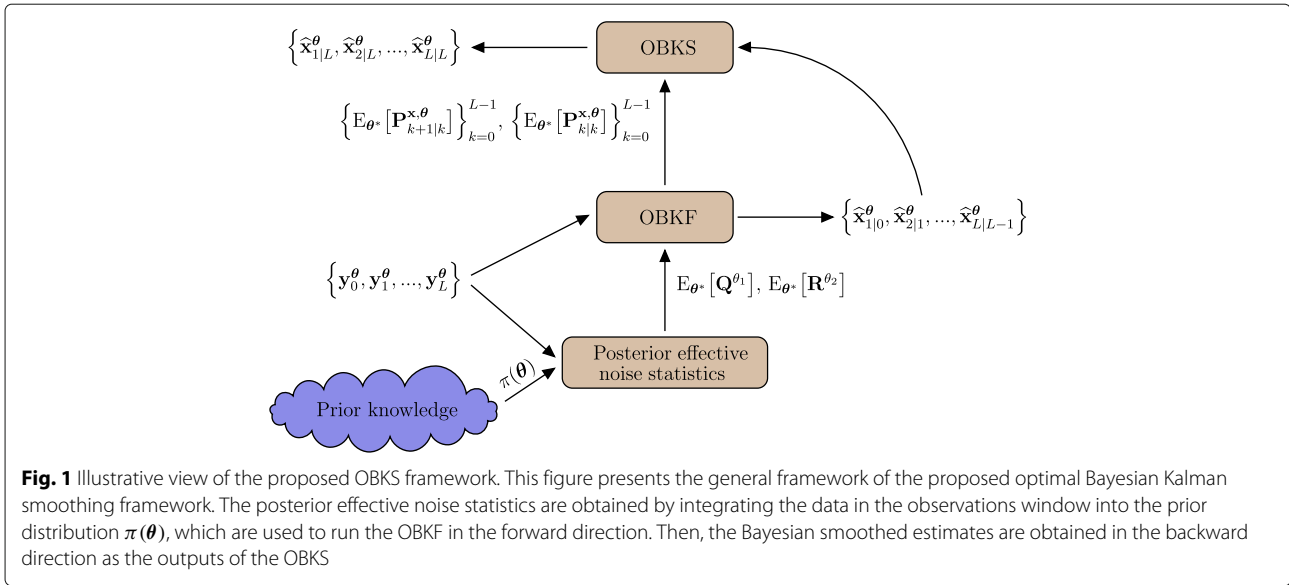
The initial values are $S_0 = 1$, $\Sigma_0 = \text{cov}[\mathbf{x}_0]$, and $\mathbf{M}_0 = \text{E}[\mathbf{x}_0]$. A pseudo-code outlining the computational steps needed for computing the likelihood function is available in Additional file 1.

The likelihood function $f(\mathcal{Y}_L^\theta | \theta)$ is needed in a Metropolis Hastings MCMC method to decide whether the generated MCMC samples should be rejected or accepted into the sequence. In this MCMC method, a sequence of samples is generated sequentially where at each step, a new sample $\theta^{(\text{new})}$ is generated based on the last accepted sample $\theta^{(\text{old})}$ according to a proposal distribution $f(\theta^{(\text{new})} | \theta^{(\text{old})})$. The new sample $\theta^{(\text{new})}$ will be either accepted or rejected based on an acceptance ratio r computed as follows:

$$r = \min \left\{ 1, \frac{f(\theta^{(\text{old})} | \theta^{(\text{new})}) f(\mathcal{Y}_L^\theta | \theta^{(\text{new})}) \pi(\theta^{(\text{new})})}{f(\theta^{(\text{new})} | \theta^{(\text{old})}) f(\mathcal{Y}_L^\theta | \theta^{(\text{old})}) \pi(\theta^{(\text{old})})} \right\}. \quad (47)$$

Note that $f(\mathcal{Y}_L^\theta | \theta^{(\text{new})})$ (and $f(\mathcal{Y}_L^\theta | \theta^{(\text{old})})$) are computed via the set of equations given in (39)–(46). The new sample $\theta^{(\text{new})}$ will be accepted into the sequence of MCMC samples with probability r . Otherwise, it will be discarded and the last sample $\theta^{(\text{old})}$ will be repeated in the sequence. When enough MCMC samples are generated, the posterior effective noise statistics $\text{E}[\theta | \mathcal{Y}_L^\theta]$ are approximated as the average of the generated samples. When a symmetric proposal distribution, i.e., $f(\theta^{(\text{old})} | \theta^{(\text{new})}) = f(\theta^{(\text{new})} | \theta^{(\text{old})})$, such as a Gaussian distribution is used in our simulations, then (47) can be further simplified. Also, more explanation and a pseudo-code on how the recursive calculations in (42)–(46) should be performed are provided in Additional file 1.

A block diagram of the proposed OBKS framework is shown in Fig. 1. As this figure shows, in an additional forward step, first the posterior effective noise statistics are computed, then these effective characteristics are used in another forward step to run the OBKF (as summarized in the forward step of Table 1), and then in the backward step, the Bayesian smoothed estimate for each state in the interval is computed as summarized in the backward step of Table 1. Note that computing the effective Kalman smoothing gain \mathbf{A}_k^\ominus in the OBKS requires $\text{E}_{\theta^*}[\mathbf{P}_{k|k}^{\mathbf{x}, \theta}]$ and $\text{E}_{\theta^*}[\mathbf{P}_{k+1|k}^{\mathbf{x}, \theta}]$, which are



the by-products of the OBKF in the forward direction. Also, all computational steps for the OBKS are summarized in Algorithm 1. The inputs are the prior distribution, the matrices that characterize the state-space model, and the observations over the observation window.

Algorithm 1 Optimal Bayesian Kalman Smoother (OBKS)

- 1: **input:** $\pi(\theta)$, Φ_k , \mathbf{H}_k , \mathbf{Q}^{θ_1} , \mathbf{R}^{θ_2} , Γ_k , $\mathcal{Y}_L^\theta = \{\mathbf{y}_0^\theta, \mathbf{y}_1^\theta, \dots, \mathbf{y}_L^\theta\}$
- 2: **output:** $\hat{\mathbf{x}}_{0|L}^\theta, \hat{\mathbf{x}}_{1|L}^\theta, \dots, \hat{\mathbf{x}}_{L|L}^\theta$
- 3: $E_{\theta^*}[\mathbf{Q}^{\theta_1}], E_{\theta^*}[\mathbf{R}^{\theta_2}] \leftarrow \text{MCMC}(\mathcal{Y}_L^\theta, \pi(\theta))$
- 4: $\hat{\mathbf{x}}_{0|0}^\theta \leftarrow E[\mathbf{x}_0]$
- 5: $\hat{\mathbf{x}}_{1|0}^\theta \leftarrow \Phi_0 E[\mathbf{x}_0]$
- 6: $E_{\theta^*}[\mathbf{P}_{0|0}^{\mathbf{x}, \theta}] \leftarrow \text{cov}[\mathbf{x}_0]$
- 7: $E_{\theta^*}[\mathbf{P}_{1|0}^{\mathbf{x}, \theta}] \leftarrow \Phi_0 \text{cov}[\mathbf{x}_0] \Phi_0^T + \Gamma_0 E_{\theta^*}[\mathbf{Q}^{\theta_1}] \text{cov}[\mathbf{x}_0] \Phi_0^T$
- 8: **for** $k = 1, \dots, L$ **do**
- 9: $\tilde{\mathbf{z}}_k^\theta \leftarrow \mathbf{y}_k^\theta - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1}^\theta$
- 10: $\mathbf{K}_k^\ominus \leftarrow E_{\theta^*}[\mathbf{P}_{k|k-1}^{\mathbf{x}, \theta}] \mathbf{H}_k^T E_{\theta^*}^{-1}[\mathbf{H}_k \mathbf{P}_{k|k-1}^{\mathbf{x}, \theta} \mathbf{H}_k^T + \mathbf{R}^{\theta_2}]$
- 11: $\hat{\mathbf{x}}_{k|k}^\theta \leftarrow \hat{\mathbf{x}}_{k|k-1}^\theta + \mathbf{K}_k^\ominus \tilde{\mathbf{z}}_k^\theta$
- 12: $\hat{\mathbf{x}}_{k+1|k}^\theta \leftarrow \Phi_k \hat{\mathbf{x}}_{k|k-1}^\theta + \Phi_k \mathbf{K}_k^\ominus \tilde{\mathbf{z}}_k^\theta$
- 13: $E_{\theta^*}[\mathbf{P}_{k|k}^{\mathbf{x}, \theta}] \leftarrow (\mathbf{I} - \mathbf{K}_k^\ominus \mathbf{H}_k) E_{\theta^*}[\mathbf{P}_{k|k-1}^{\mathbf{x}, \theta}]$
- 14: $E_{\theta^*}[\mathbf{P}_{k+1|k}^{\mathbf{x}, \theta}] \leftarrow \Phi_k (\mathbf{I} - \mathbf{K}_k^\ominus \mathbf{H}_k) E_{\theta^*}[\mathbf{P}_{k|k-1}^{\mathbf{x}, \theta}] \Phi_k^T + \Gamma_k E_{\theta^*}[\mathbf{Q}^{\theta_1}] \Gamma_k^T$
- 15: **for** $k = L-1, L-2, \dots, 0$ **do**
- 16: $\mathbf{A}_k^\ominus \leftarrow E_{\theta^*}[\mathbf{P}_{k|k}^{\mathbf{x}, \theta}] \Phi_k^T E_{\theta^*}^{-1}[\mathbf{P}_{k+1|k}^{\mathbf{x}, \theta}]$
- 17: $\hat{\mathbf{x}}_{k|L}^\theta \leftarrow \hat{\mathbf{x}}_{k+1|L}^\theta + \mathbf{A}_k^\ominus (\hat{\mathbf{x}}_{k+1|L}^\theta - \hat{\mathbf{x}}_{k+1|k}^\theta)$
- 18: $E_{\theta^*}[\mathbf{P}_{k|L}^{\mathbf{x}, \theta}] \leftarrow E_{\theta^*}[\mathbf{P}_{k|k}^{\mathbf{x}, \theta}] + \mathbf{A}_k^\ominus (E_{\theta^*}[\mathbf{P}_{k+1|L}^{\mathbf{x}, \theta}] - E_{\theta^*}[\mathbf{P}_{k+1|k}^{\mathbf{x}, \theta}]) (\mathbf{A}_k^\ominus)^T$
- return** $\hat{\mathbf{x}}_{k|L}^\theta$

The outputs are the Bayesian smoothed estimates of the unknown states over the window obtained by the OBKS. There are four main steps in this algorithm. First, in line 3, the posterior effective noise statistics are estimated using the MCMC approach, as explained in Section 2.4, which are later used in the OBKF. Then, we need to initialize the OBKF as outlined in lines 4–7. Lines 8–14 are devoted to the OBKF, which is run in the forward direction. Finally, lines 15–18 show how the Bayesian smoothed estimates are obtained.

Also, in order to study the computational complexity of the proposed OBKS, since the proposed recursive structure is completely similar to that of the ordinary Kalman smoother except using the posterior effective noise statistics, which are approximated using MCMC samples, we only need to analyze the complexity of the MCMC step. In the MCMC step, in order to obtain a sequence of MCMC samples, the likelihood function in (36) should be computed for each generated MCMC sample by iterating the equations given in (42)–(46) from $k = 0$ to $k = L - 1$. Therefore, the dimensions of the state vector \mathbf{x} and the observation vector \mathbf{y} , the size of the window L , and the number of generated MCMC samples can affect the complexity. We should point out that some matrix calculations such as inversions, multiplications, and determinants might need to be performed only one time for each generated MCMC sample. For example, when the process noise covariance matrix is known and the system is stationary, it is enough to compute $\Phi_k^T (\tilde{\mathbf{Q}}_k^{\theta_1})^{-1}$ once and then use it for the rest of the calculations.

3 Simulation results and discussion

To study the performance of different smoothing approaches, we need the smoothing error covariance

matrix $\mathbf{P}_{k|L}^{\theta',\theta}$ that characterizes the performance of the Kalman smoother designed by the assumption of $\theta' = [\theta'_1, \theta'_2]$ when applied to a model with the actual noise parameters $\theta = [\theta_1, \theta_2]$, which is given by [35]

$$\begin{aligned} \mathbf{P}_{k|L}^{\theta',\theta} &= \bar{\mathbf{A}}_k^{\theta'} \mathbf{P}_{k|k}^{\theta',\theta} (\bar{\mathbf{A}}_k^{\theta'})^T + \mathbf{A}_k^{\theta'} (\tilde{\mathbf{Q}}_k^{\theta_1} + \mathbf{P}_{k+1|L}^{\theta',\theta}) (\mathbf{A}_k^{\theta'})^T \\ &\quad + \bar{\mathbf{A}}_k^{\theta'} \mathbf{P}_{k|k}^{\theta',\theta} \Phi_k^T (\mathbf{I} - \mathbf{D}_k^{\theta'})^T (\mathbf{A}_k^{\theta'})^T + \mathbf{A}_k^{\theta'} (\mathbf{I} - \mathbf{D}_k^{\theta'}) \Phi_k \mathbf{P}_{k|k}^{\theta',\theta} (\bar{\mathbf{A}}_k^{\theta'})^T \\ &\quad - \mathbf{A}_k^{\theta'} ((\tilde{\mathbf{Q}}_k^{\theta_1})^T (\mathbf{I} - \mathbf{L}_k^{\theta'})^T + (\mathbf{I} - \mathbf{L}_k^{\theta'}) \tilde{\mathbf{Q}}_k^{\theta_1}) (\mathbf{A}_k^{\theta'})^T, \end{aligned} \quad (48)$$

where $\bar{\mathbf{A}}_k^{\theta'} = \mathbf{I} - \mathbf{A}_k^{\theta'} \Phi_k$, $\mathbf{A}_k^{\theta'}$, and $\mathbf{K}_k^{\theta'}$ are the Kalman smoothing gain and Kalman gain matrices of the Kalman smoother and filter designed relative to θ' , respectively, and matrices $\mathbf{D}_k^{\theta'}$ and $\mathbf{L}_k^{\theta'}$ can be found recursively via

$$\begin{aligned} \mathbf{D}_k^{\theta'} &= \mathbf{A}_{k+1}^{\theta'} \mathbf{D}_{k+1}^{\theta'} \Phi_{k+1} (\mathbf{I} - \mathbf{K}_{k+1}^{\theta'} \mathbf{H}_{k+1}) + \mathbf{K}_{k+1}^{\theta'} \mathbf{H}_{k+1}, \\ \mathbf{L}_k^{\theta'} &= \mathbf{D}_k^{\theta'}. \end{aligned} \quad (49)$$

The initial conditions for these two matrices are $\mathbf{D}_{L-1}^{\theta'} = \mathbf{K}_L^{\theta'} \mathbf{H}_L$ and $\mathbf{L}_{L-1}^{\theta'} = \mathbf{K}_{L-1}^{\theta'} \mathbf{H}_{L-1}$.

We compare the OBKS with the steady-state minimax Kalman smoother, intrinsically Bayesian robust Kalman smoother (IBR-KS), and the model-specific Kalman smoothers designed relative to all possible values of θ . The minimax Kalman smoother has the best worst-case performance among all possible model-specific smoothers and is defined by

$$\theta_{\text{mm}} = \arg \min_{\theta' \in \Theta} \max_{\theta \in \Theta} \text{Tr} \left\{ \mathbf{P}_{\frac{L}{2}|L}^{\theta',\theta} \right\}. \quad (50)$$

Note that we focus on the state at time $L/2$, which is in the middle of the observation window, as the steady-state performance of a smoother occurs for the middle points in an observation window. The IBR-KS is similar to the OBKS except that the optimization is relative to the prior distribution $\pi(\theta)$. To design an IBR-KS, one can use the OBKS equations in Table 1 with expectations relative to the prior distribution rather than the posterior distribution. Therefore, for the IBR-KS the MCMC step is not needed. The IBR-KS approach provides optimal smoothing performance relative to the prior distribution.

3.1 Target tracking example

Let the dynamic of a vehicle at each time step be determined by the state vector $\mathbf{x}_k = [p_x \ v_x \ p_y \ v_y]^T$, where p_x , v_x , p_y , and v_y are the horizontal position, horizontal velocity, vertical position, and vertical velocity, respectively. If the vehicle possesses a constant speed and the measurements are made with intervals τ , then a state-space model with

the following matrices can characterize the dynamic of the vehicle at each time step [36–38]:

$$\Phi_k = \begin{bmatrix} 1 & \tau & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H}_k = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \Gamma_k = \mathbf{I}.$$

The covariance matrices of the process noise and observation noise are

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

where q determines the process noise intensity. In the simulations, we assume that the measurement interval is $\tau = 1$ and the initial conditions are $E[\mathbf{x}_0] = [100 \ 10 \ 30 \ -10]^T$ and $\text{cov}[\mathbf{x}_0] = \text{diag}[25 \ 2 \ 25 \ 2]$.

It is worth mentioning that when the process noise transition matrix Γ_k is not an identity matrix, since the effect of Γ_k on the Kalman equations is only through the covariance matrix of the process noise, a state-space model with the process noise covariance matrix \mathbf{Q} and the process noise transition matrix Γ_k is equivalent to a state-space model with the process noise covariance matrix $\mathbf{Q}_k^{\text{eq}} = \Gamma_k \mathbf{Q} \Gamma_k^T$ and the process noise transition matrix $\Gamma_k^{\text{eq}} = \mathbf{I}$. With this in mind, we can consider the simulation results of the above target tracking example for an equivalent state-space model with the process noise transition matrix

$$\Gamma_k^{\text{eq}} = \begin{bmatrix} -0.8817 & 0 & 0 & 0.4719 \\ 0.4719 & 0 & 0 & 0.8817 \\ 0 & -0.8817 & 0.4719 & 0 \\ 0 & 0.4719 & 0.8817 & 0 \end{bmatrix},$$

and the covariance matrix $\mathbf{Q}' = q \times \text{diag}[0.0657 \ 0.0657 \ 1.2676 \ 1.2676]$, where \mathbf{Q}' is the diagonal matrix of the eigenvalues and Γ_k^{eq} is matrix of the eigenvectors for \mathbf{Q} , i.e., $\mathbf{Q} = \Gamma_k^{\text{eq}} \mathbf{Q}' (\Gamma_k^{\text{eq}})^T$ is the eigen-decomposition of \mathbf{Q} . Therefore, the simulations for the target tracking example can also be regarded for the case that the process noise transition matrix is not an identity matrix.

We set q to 2 and assume that the diagonal element r of the observation noise covariance matrix is unknown and represented by a uniform random variable θ over $[0.25, 5]$. In previous work, we have used the inverse-Wishart distribution as a prior for the covariance matrices [20] and could have done so here; however, for computational reasons, we have limited our example to priors governing q and r . Let the observation window be of length $L = 15$. For the MCMC step, we generate 10,000 MCMC samples and use a Gaussian distribution as the proposal distribution with the mean being the last accepted MCMC sample and variance being equal to 4. First, we study the average performance (MSE) over the uncertainty class for

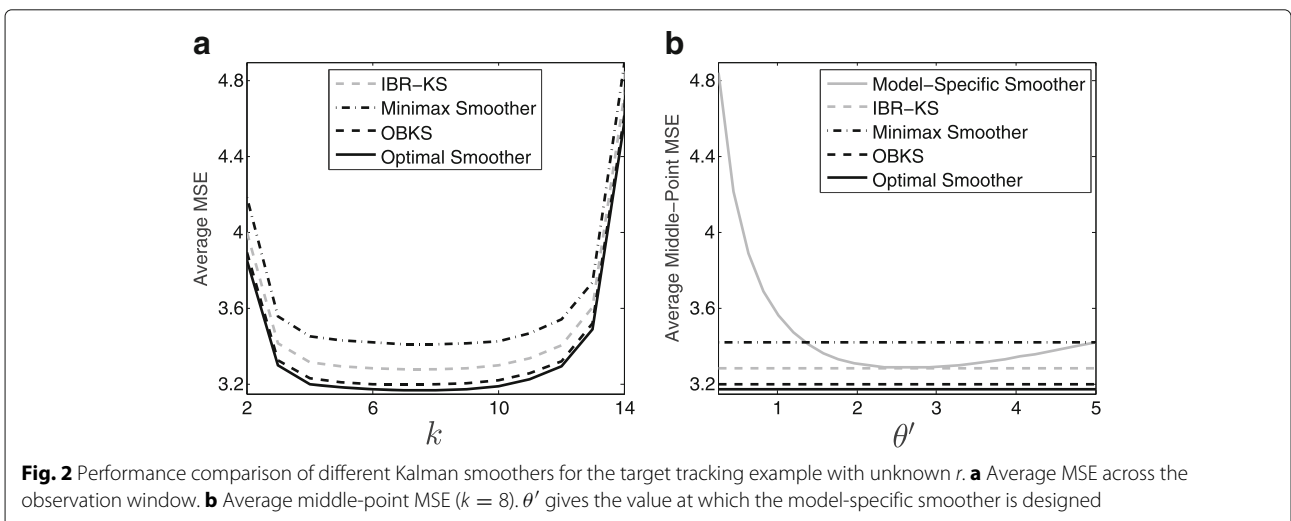
various Kalman smoothing approaches. To do so, for a given sequence of observations \mathcal{Y}_L^θ , we find the error of each Kalman smoothing approach for each k by first computing the error covariance matrix using (48), where θ' is replaced by $E[\theta|\mathcal{Y}_L^\theta]$, $E[\theta]$, and θ_{mm} for the OBKS, IBR-KS, and the minimax Kalman smoother, respectively, and then, the MSE at time k is obtained by finding the sum of diagonal elements of the error covariance matrix. The MSE for the optimal Kalman smoother designed relative to the actual θ value is obtained according to $\mathbf{P}_{k|L}$ in Table 1. The reported average MSE is obtained over 30 different assumed values of θ and 10 different observation sequences for each value (300 simulations in total). Figure 2a presents the average MSE across the observation window obtained for each smoothing scheme. As can be seen, OBKS outperforms IBR and minimax approaches and its performance is close to the average of the optimal MSEs obtained by the optimal smoothers. Figure 2b shows the average MSE for the middle state ($k = 8$) in the observation window. In addition, this figure presents the average MSE of each model-specific Kalman smoother designed relative to value θ' . Note that the difference between the definitions of the optimal smoother and the model-specific smoother is that the optimal smoother is designed relative to θ and always applied to model θ , but the model-specific smoother is designed relative to θ' and then applied to the model θ . This figure suggests the better performance of the OBKS compared to the IBR, minimax, and model-specific Kalman smoothers.

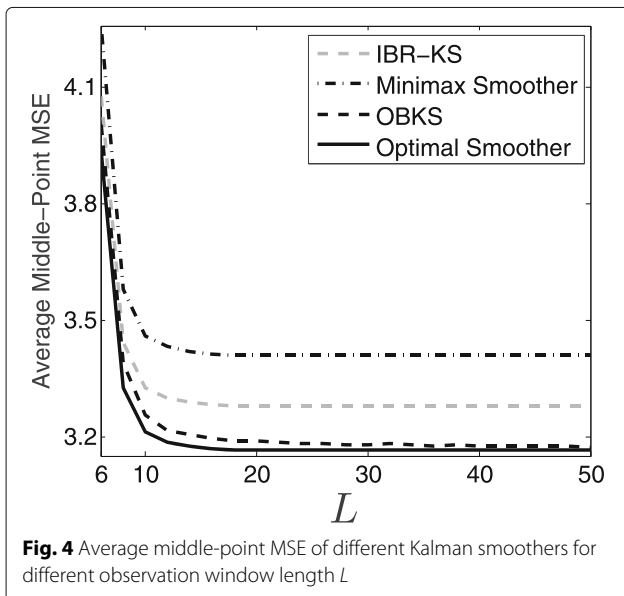
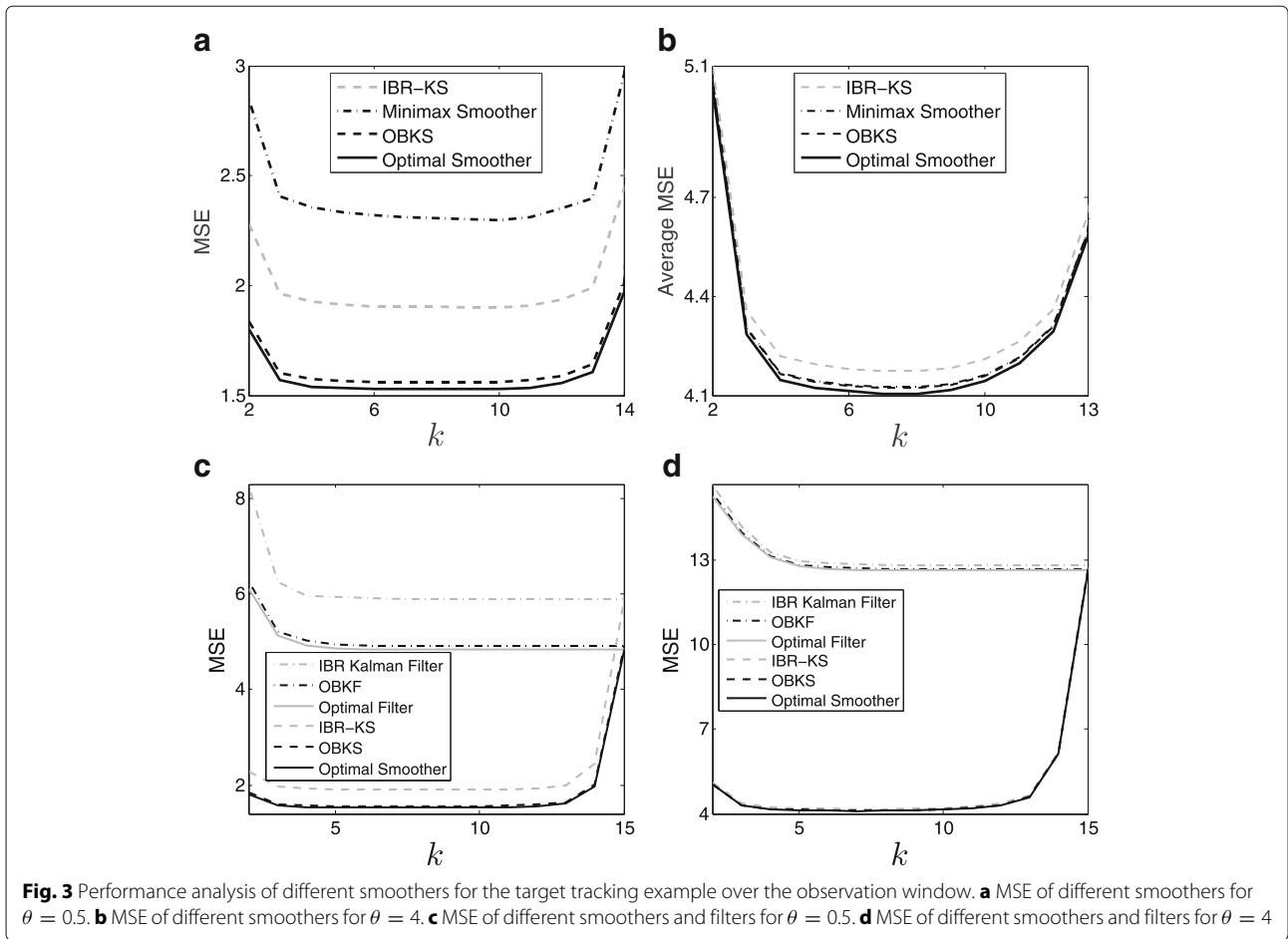
Figure 3a, b illustrate the performances of different Kalman smoothers for the observation noise variances $\theta = 0.5$ and $\theta = 4$, respectively. Although θ is fixed, since the OBKS performance depends on the generated observations, we report the average MSE taken over 300

different generated observation sequences \mathcal{Y}_L^θ . It can be seen that the OBKS has its performance close to the optimal Kalman smoothers and much better compared to other robust Kalman smoothers. Since the IBR approach is optimal on average relative to the prior distribution, not for each possible model within the class, it is not guaranteed that the IBR-KS performs well for each model, an example being $\theta = 4$ where minimax outperforms the IBR approach. However, even for models for which the IBR approach does not perform well, OBKS still gives promising results.

In Fig. 3c, d, in addition to the MSE values for different smoothers, we also present the MSEs of various Kalman filters for each time step k . For different filters, we compute $\mathbf{P}_{k|k}^{\theta',\theta}$, as derived in [16], and report $\text{Tr}\{\mathbf{P}_{k|k}^{\theta',\theta}\}$, where θ' is replaced by θ_{mm} , $E[\theta]$, and $E[\theta|\mathcal{Y}_L^\theta]$ for the minimax, IBR, and OBKF approaches, respectively. Since the error covariance matrix of the filter is used as the initial value for the error covariance matrix of the corresponding smoother, for $k = 15$, the MSE of each smoother equals that of its corresponding filter. As expected, for other time indices k within the observation window, the MSE of the smoother is always lower than that of the filter. Note that due to the range of the y -axis in Fig. 3c, d, the MSEs of various smoothing approaches might not be distinguishable. The difference between the performances of different smoothers is visible in Fig. 3a, b.

Next, we analyze the smoothing performance for different numbers of observations (observation window length) L . Figure 4 presents the average MSE of different Kalman smoothers at time step $L/2$ (middle state) as L changes from 6 to 50. The average MSE for each L is obtained in the same way as in Fig. 2. We can observe





that when the number of observations is small, the performances of OBKS and IBR-KS are close because the expectation of θ relative to the posterior distribution is close to the expectation relative to the prior distribution; however, as the number of observations increases, the average MSE of the OBKS gets closer to that of the optimal smoothers because the expectation of the unknown parameter tends to the true parameter value. Moreover, both the OBKS and IBR-KS always outperform the minimax Kalman smoother in terms of average MSE.

In the next set of simulations, we consider the case that both the process noise parameter q and the observation noise parameter r are unknown, being denoted by uniform random variables θ_1 and θ_2 over intervals $[3, 5]$ and $[0.25, 5]$, respectively. Regarding the MCMC step, we use a multivariate Gaussian distribution with the mean vector being the vector of last accepted samples for θ_1 and θ_2 and a diagonal covariance matrix whose diagonal elements are $\text{diag}[1 \ 1.5]$. Analogous to the previous set of simulations, we compare the performance of the OBKS with other smoothing approaches in terms of the average MSE over 200 different assumed true values for θ_1 and θ_2

and 10 different sets of observations for each pair of true values (2000 simulations in total) in Fig. 5. As shown in the figure, the OBKS outperforms other robust smoothers and performs closely to the optimal smoother designed relative to the underlying true model.

In Fig. 6, we analyze the average MSE at $k = 8$, the middle point in the observation window. The white surface represents the average middle-point MSE for each model-specific Kalman smoother designed relative to the process noise parameter θ'_1 and observation noise parameter θ'_2 . We also show the average middle-point MSEs for the IBR-KS, minimax smoother, and OBKS, and the average of the optimal middle-point MSEs obtained by the optimal smoothers by constant planes. This figure suggests that compared to other robust smoothers, the OBKS achieves the closest average middle-point MSE to that of the optimal smoothers.

Figure 7 shows the performance of different smoothing approaches for two specific state-space models corresponding to certain values of θ_1 and θ_2 . Figure 7a, c correspond to $\theta_1 = 4.5$ and $\theta_2 = 1$, respectively, and Fig. 7b, d correspond to $\theta_1 = 3.2$ and $\theta_2 = 4$, respectively. The figures in the first row report the MSE of different smoothers for each time row instance within the observation window. For both state-space models, we see the promising performance of the OBKS. In addition to the MSEs of different smoothers, the second row gives the MSEs of different Kalman filters. The MSE of each smoother is initialized by the MSE of the corresponding filter, and then, it decreases as we proceed in the backward direction. The difference between the performances

of different smoothers is visible in Fig. 7a, b, which focus on a shorter range for the smoothing performance.

Figure 8 studies the effect of the size of the observation window on the performances of different smoothing approaches. We vary the size of window L from 6 to 50 and report the average middle-point MSEs of various smoothers for each L . When L is small, the performances of the OBKS and IBR-KS are close but as L increases, the performance of the OBKS tends to that of the optimal smoother. This is because the posterior effective noise statistics converge to the underlying true values as the number of observations increases.

In Fig. 9, we analyze the complexity of the proposed OBKS framework and how its runtime changes with the size of the window and the number of MCMC samples. We consider the target tracking example when the size of the window changes from 4 to 50 and the number S of generated MCMC samples is 5000, 10,000, and 15,000. Computations were performed on a machine with 16 GB RAM and Intel® Core™ i7 2.5 GHz CPU. As can be seen, the run time tends to grow linearly with L and S . In our simulations, we set the number of samples in the MCMC step to 10,000 to obtain acceptable estimates at tolerable computational complexity.

Also, in Fig. 10, we study the effect of the number of MCMC samples, used to compute the posterior effective noise statistics, on the OBKS performance. Similar to Fig. 3a, we assume that the observation noise variance θ is unknown and its true value is 0.5. We consider three different observation window sizes at $L = 10, 15, 20$ and vary the number of MCMC samples from 100 to 10,000. For each L and number of MCMC samples, we report the middle-point MSE (MSE at $k = L/2$) over 300 different observation sequences generated based on the underlying true state-space model. As can be seen, as the number of MCMC samples increases (especially when the number of MCMC samples is not large enough), the OBKS performance gets better because more accurate posterior effective noise statistics can be obtained via more MCMC samples. However, after collecting enough MCMC samples, the performance of the OBKS converges, and further increase of MCMC samples has little additional performance improvement. In our simulations throughout the paper, we used 10,000 MCMC samples.

3.2 Gene regulatory network inference

In this section, we apply the proposed OBKS framework to gene regulatory network (GRN) inference. GRNs are used as a platform to characterize the relationship between genes and play a major role in drug design. Numerous GRN inference methods have been proposed in the literature, one based on Kalman filtering [39] in which the inference problem is formulated as a state-space problem where the parameters to be inferred are regarded as

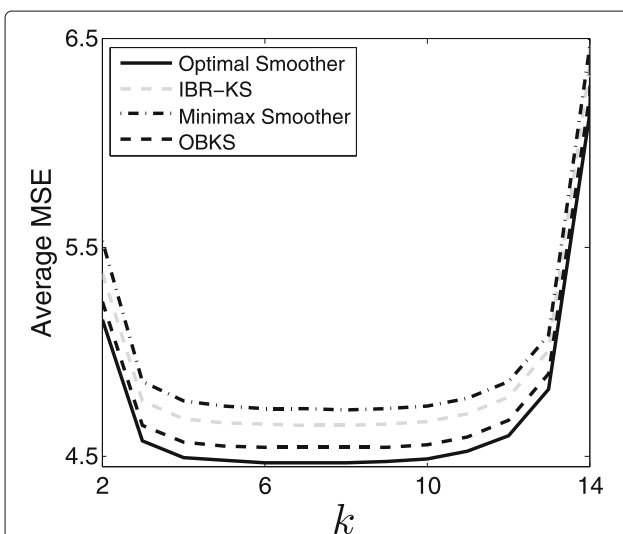


Fig. 5 Performance analysis over the observation window for the target tracking example with unknown r and q . Average MSE of different smoothers for each k over the observation window when both q and r are unknown

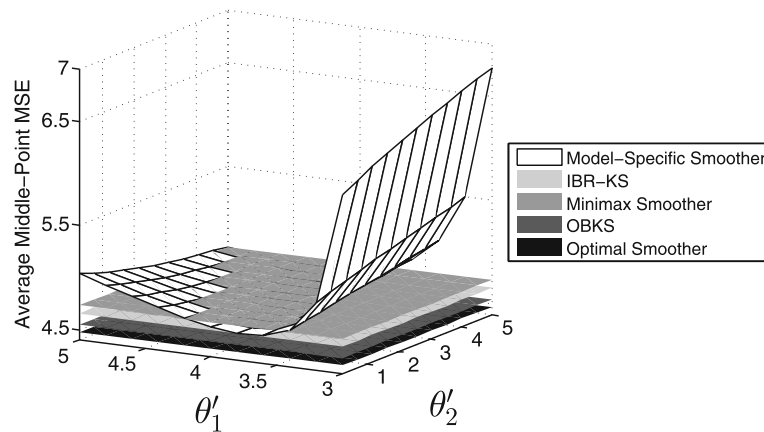


Fig. 6 Average middle-point MSE of different Kalman smoothers for the target tracking example with unknown r and q . The average performance of each model-specific Kalman smoother corresponding to the noise parameters θ_1' and θ_2' is shown. The average MSE for the IBR-KS, minimax, OBKS, and the average of optimal MSEs are shown as constant planes

hidden states. Since acquiring exact knowledge of noise statistics is highly difficult due to the complexity of biological systems and other practical limitations, it is prudent to utilize a robust Kalman approach for inference. We focus on the continuous nonlinear ordinary differential

equation model [39], where the value of each gene g_i , $1 \leq i \leq n$, n being the total number of genes in the network, is characterized as

$$\dot{g}_i = \eta_i(g_1, \dots, g_n) + v_i,$$

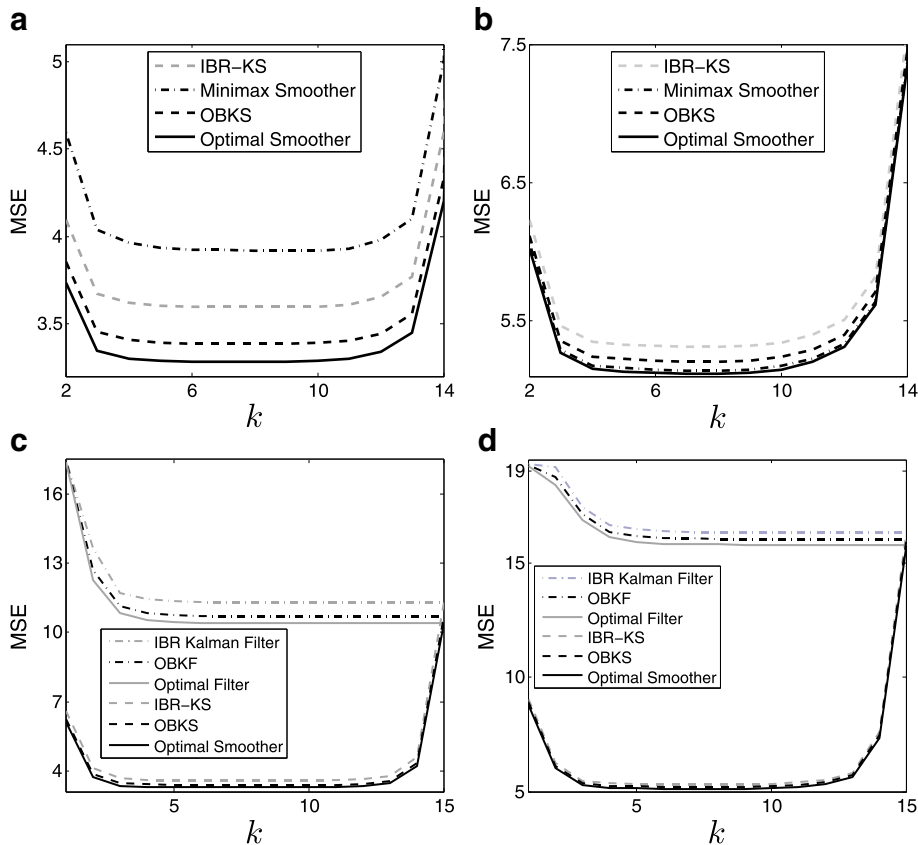


Fig. 7 Performance of different smoothing approaches when applied to specific state-space models for the target tracking example with unknown r and q . **a** MSE of different smoothers for $\theta_1 = 4.5$ and $\theta_2 = 1$. **b** MSE of different smoothers for $\theta_1 = 3.2$ and $\theta_2 = 4$. **c** MSE of different smoothers and filters for $\theta_1 = 4.5$ and $\theta_2 = 1$. **d** MSE of different smoothers and filters for $\theta_1 = 3.2$ and $\theta_2 = 4$

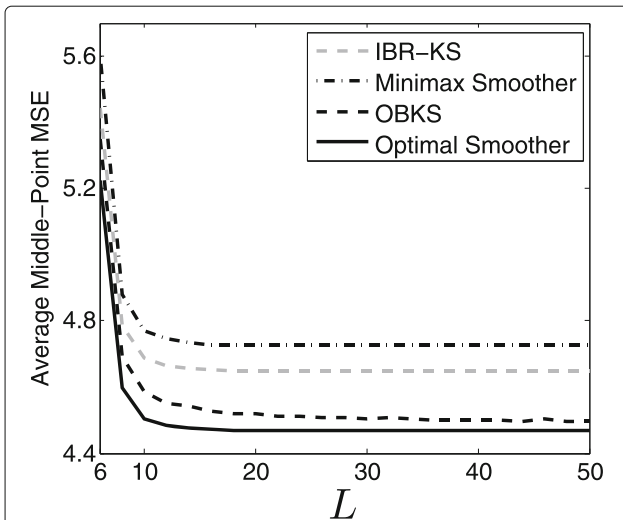


Fig. 8 Effect of the size of observations on the smoother performance in the target tracking example with unknown r and q

where \dot{g}_i , v_i , and $\eta_i(\cdot)$ are the derivative of the gene-expression value relative to the time variable, the external noise, and the regulatory function, respectively. The regulatory function η_i is a linear combination of some nonlinear terms [39]:

$$\eta_i(g_1, \dots, g_n) = \sum_{j=1}^{N_i} [(\alpha_{ij} + u_{ij})\Omega_{ij}(g_1, \dots, g_n)],$$

where N_i is the number of nonlinear terms in η_i , $\Omega_{ij}(\cdot)$ is the j th nonlinear term in η_i with corresponding coefficient α_{ij} and parameter noise u_{ij} .

The inference problem for this model involves estimating the values of coefficients α_{ij} from time series data, generated from the underlying true GRN model. To estimate unknown coefficients from data, following [16, 39],

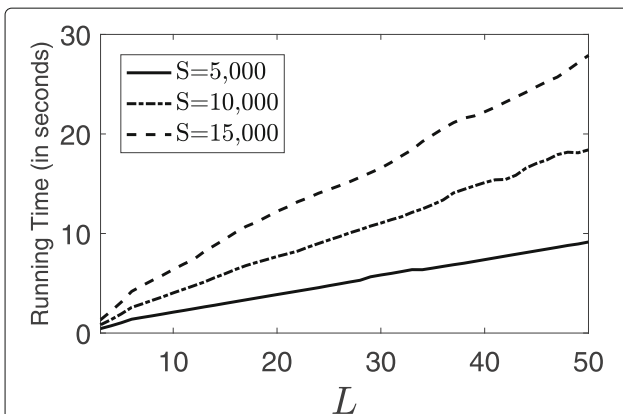


Fig. 9 Processing time required for implementing the OBKS for the target tracking example relative to the size of the window L and the number of MCMC samples S

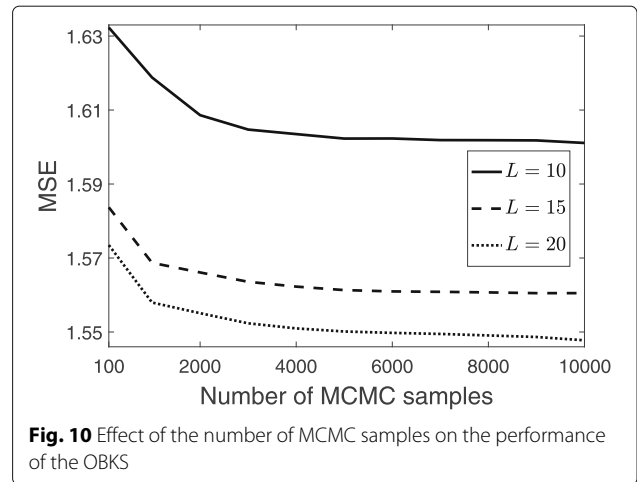
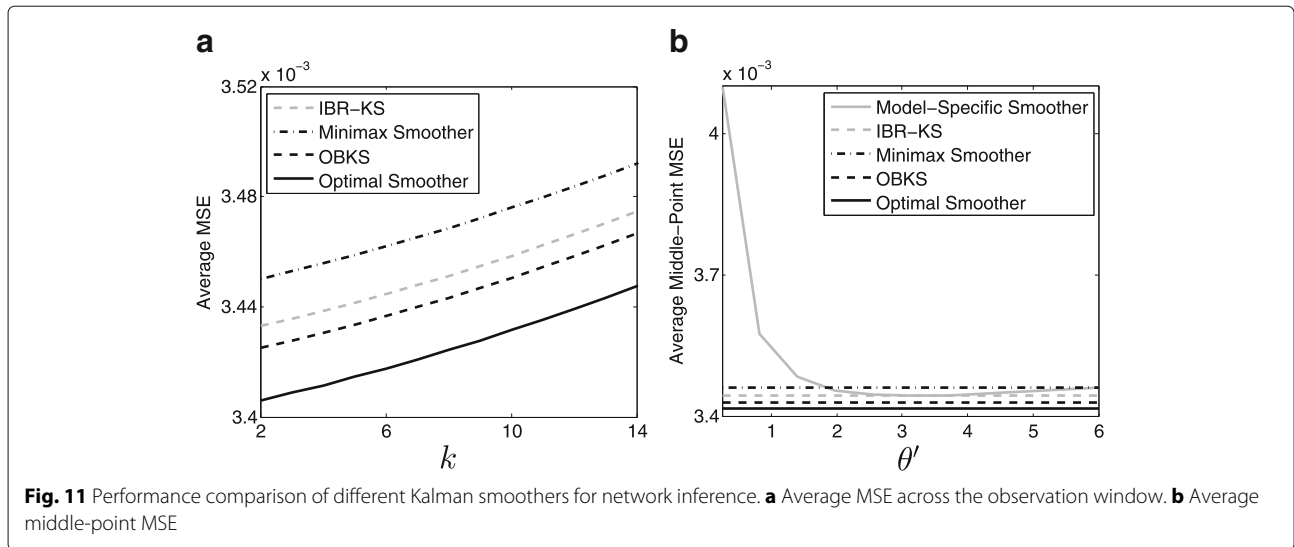


Fig. 10 Effect of the number of MCMC samples on the performance of the OBKS

we build a state-space model with vectors formed by stacking coefficients α_{ij} , parameter noise u_{ij} , and external noise v_i in place of the state vector \mathbf{x}_k , process noise vector \mathbf{u}_k , and observation noise vector \mathbf{v}_k , respectively. In the state-space model, we have $\Phi_k = \mathbf{I}$ and $\Gamma_k = \mathbf{I}$. The observation vector \mathbf{y}_k and the observation transition matrix \mathbf{H}_k are formed using gene-expression values g_i and the nonlinear terms Ω_{ij} , respectively. More details on constructing the state-space model for this inference problem can be found in [16, 39]. In this paper, we work out the inference problem for the yeast cell cycle network [39], which has $n = 12$ genes and 54 coefficients to be inferred. For this network, the state vector is of size 54 and the observation vector is of size 12. To evaluate the performance of the OBKS for network inference, we use the synthetic time series data generated according to the regulatory equations given in [39].

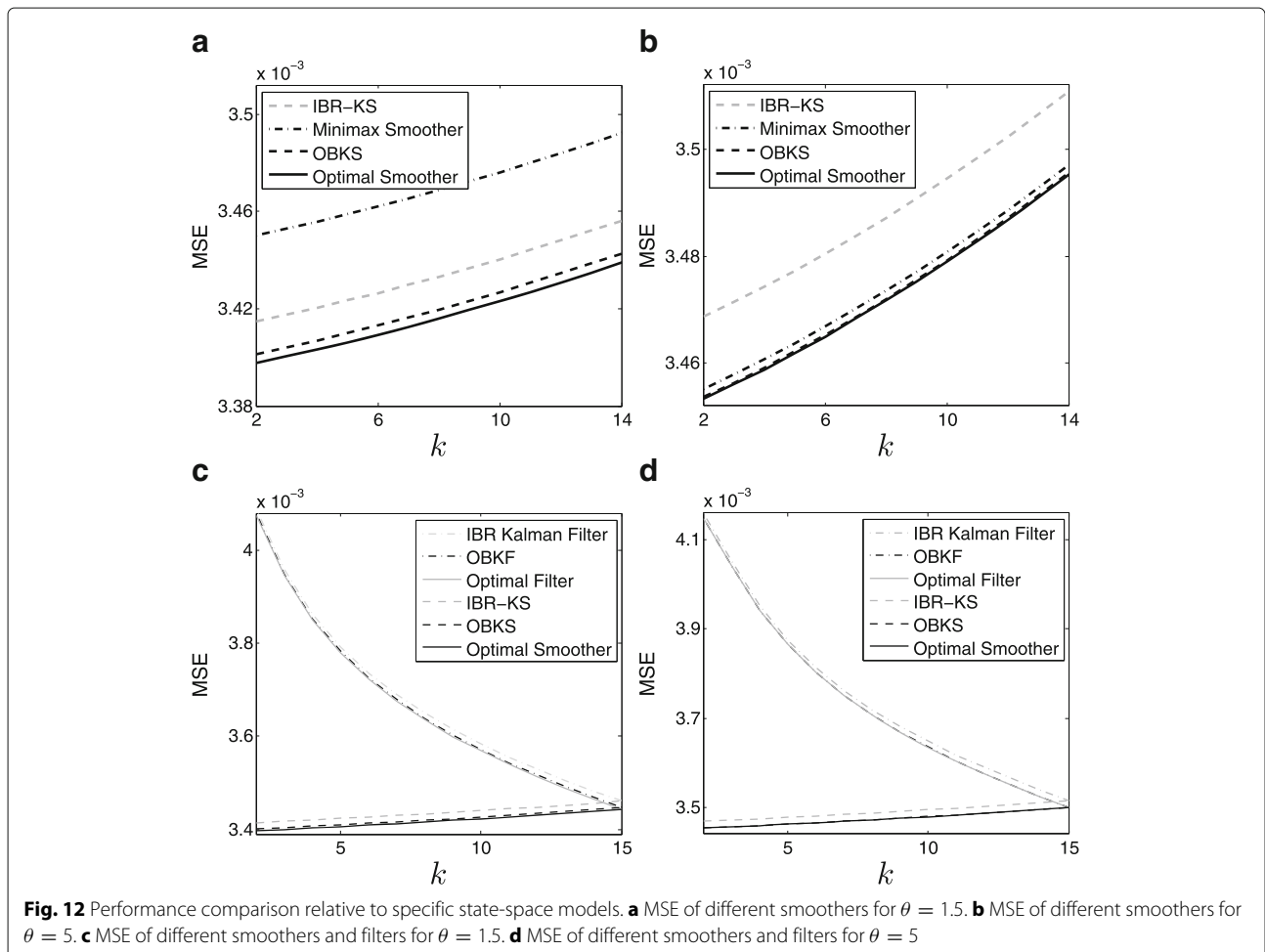
In our simulations, we assume $\mathbf{Q} = 10^{-7} \times \mathbf{I}$ and $\mathbf{R} = \theta \times \mathbf{I}$ in which θ is unknown and belongs to $[0.25, 6]$. Let the initial conditions be $E[\mathbf{x}_0] = \mathbf{0}_{54 \times 1}$ and $\text{cov}[\mathbf{x}_0] = 0.001 \times \mathbf{I}$. We set the observation window length L to 15. For MCMC calculations, we use the same setting as the first set of simulations for the target tracking example. First we analyze the average performance over the uncertainty class. In order to report the average MSE for each smoothing scheme, we take the average of the MSEs obtained over 30 different assumed true values of θ and 20 different observation sequences for each assumed true value (600 simulations). In Fig. 11a, we compare the average MSEs of different Kalman smoothers for each time index k within the observation window. The average MSE from the OBKS is lower compared to those of the IBR-KS and minimax smoothers. Fig. 11b presents the average middle-point MSE (for $k = 8$) for each model-specific Kalman smoother designed relative to value θ' for the noise parameter and also those for the IBR-KS, minimax, and the OBKS approaches are shown. We also



show the average of the optimal MSEs obtained by the optimal smoothers. This figure verifies the promising performance of the OBKS approach.

Figure 12 illustrates the performance of different approaches for two specific state-space models corresponding to $\theta = 1.5$ and $\theta = 5$. For each assumed true value, the results are averaged over 200 different observation sequences generated based on the underlying true value of θ . In Fig. 12a and b, we only show the performance of different smoothers, but in Fig. 12c and d, we

ponding to $\theta = 1.5$ and $\theta = 5$. For each assumed true value, the results are averaged over 200 different observation sequences generated based on the underlying true value of θ . In Fig. 12a and b, we only show the performance of different smoothers, but in Fig. 12c and d, we



show the performances for both the smoothing and filtering schemes. We can see that the OBKS performs much better compared to other robust approaches.

The effect of the observation window size on the performance of Kalman smoother-based network inference is studied in Fig. 13. For each window size L , we report the average MSE for $k = L/2$ corresponding to each Kalman smoothing strategy. For each L , the average MSEs are obtained in the same way as Fig. 11. As shown in the figure, the performance of the OBKS gets closer to that of the optimal smoother for larger L , which is what we expect from the OBKS as the posterior effective noise statistics, relative to which the OBKS is designed, eventually converge to the underlying true values.

4 Conclusions

We proposed an optimal Bayesian Kalman smoothing framework that provides the optimal smoothing performance relative to the posterior distribution of the unknown noise parameters. Thanks to the effective Kalman smoothing gain that is applied to the posterior distribution, the structure of the proposed OBKS is analogous to that of the classical Kalman smoother. In the absence of the prior update step via factor graph, one can employ the IBR Kalman smoother to obtain the optimality relative to the prior distribution. The optimal Bayesian smoothing framework can play a major role in applications where data are rare or expensive, such as in genomics.

There are a few avenues of research in which our future work can proceed. One future direction is to address prior

construction for the proposed OBKS framework, which involves optimizing the prior distribution such that it can reflect the available prior knowledge as perfectly as possible. For example, this has been done for genomic classification by utilizing gene signaling pathway knowledge to optimize prior distribution parameters [40]. Another avenue is to extend the OBKS framework to other state-space models in which noise is not white or the state-space model is not linear, which takes the OBKS to the realm of extended Kalman filters.

Additional file

Additional file 1: Supplementary materials. This is a file in PDF format that provides a pseudo-code outlining the required steps for the factor graph-based approach for computing the likelihood function. (PDF 253 kb)

Funding

This work was funded in part by Award CCF-1553281 from the National Science Foundation.

Availability of data and materials

Data and MATLAB source code are available from the corresponding author upon request.

Authors' contributions

RD conceived the method, developed the algorithm, performed the simulations, analyzed the results, and wrote the first draft. XQ analyzed the results and edited the manuscript. ERD conceived the method, oversaw the project, analyzed the results, and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

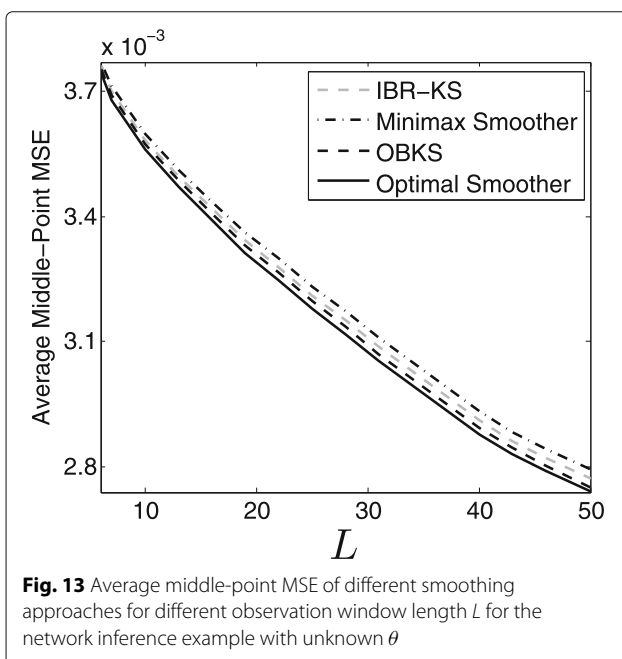
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 21 January 2018 Accepted: 13 August 2018

Published online: 06 September 2018

References

1. R.E. Kalman, A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**(1), 35–45 (1960)
2. C.K. Chui, G. Chen, et al., *Kalman filtering*. (Springer, Cham, 2017)
3. C.C. Drovandi, J.M. McGree, A.N. Pettitt, A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design. *J. Comput. Graph. Stat.* **23**(1), 3–24 (2014)
4. N. Chopin, P.E. Jacob, O. Papaspiliopoulos, SMC2: an efficient algorithm for sequential analysis of state space models. *J. R. Stat. Soc. Ser. B Stat Methodol.* **75**(3), 397–426 (2013)
5. D. Crisan, J. Miguez, et al., Nested particle filters for online parameter estimation in discrete-time state-space Markov models. *Bernoulli.* **24**(4A), 3039–3086 (2018)
6. L. Martino, J. Read, V. Elvira, F. Louzada, Cooperative parallel particle filters for online model selection and applications to urban mobility. *Digit. Signal Proc.* **60**, 172–185 (2017)
7. I. Urteaga, M.F. Bugallo, P.M. Djurić, in *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. Sequential Monte Carlo methods under model uncertainty (IEEE, 2016), pp. 1–5
8. K.A. Myers, B.D. Tapley, Adaptive sequential estimation with unknown noise statistics. *IEEE Trans. Autom. Control.* **21**(4), 520–523 (1976)
9. R.K. Mehra, On the identification of variances and adaptive Kalman filtering. *IEEE Trans. Autom. Control.* **15**(2), 175–184 (1970)



10. H.V. Poor, On robust Wiener filtering. *IEEE Trans. Autom. Control.* **25**(3), 531–536 (1980)
11. S. Verdu, H. Poor, On minimax robustness: a general approach and applications. *IEEE Trans. Infor. Theory.* **30**(2), 328–340 (1984)
12. V. Poor, D.P. Looze, Minimax state estimation for linear stochastic systems with noise uncertainty. *IEEE Trans. Autom. Control.* **26**(4), 902–906 (1981)
13. A.M. Grigoryan, E.R. Dougherty, Design and analysis of robust binary filters in the context of a prior distribution for the states of nature. *Math. Imag. Vision.* **11**(3), 239–254 (1999)
14. A.M. Grigoryan, E.R. Dougherty, Bayesian robust optimal linear filters. *Signal Process.* **81**(12), 2503–2521 (2001)
15. L. Dalton, E. Dougherty, Intrinsically optimal Bayesian robust filtering. *IEEE Trans. Signal Process.* **62**(3), 657–670 (2014)
16. R. Dehghannasiri, M.S. Esfahani, E.R. Dougherty, Intrinsically Bayesian robust Kalman filter: an innovation process approach. *IEEE Trans. Signal Process.* **65**(10), 2531–2546 (2017)
17. X. Qian, E. Dougherty, Bayesian regression with network prior: optimal Bayesian filtering perspective. *IEEE Trans. Signal Process.* **64**(23), 6243 (2016)
18. R. Dehghannasiri, M.S. Esfahani, X. Qian, E.R. Dougherty, Optimal Bayesian Kalman filtering with prior update. *IEEE Trans. Signal Process.* **66**(8), 1982–1996 (2018)
19. L.A. Dalton, E.R. Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework—part I: discrete and Gaussian models. *Pattern Recogn.* **46**(5), 1301–1314 (2013)
20. R. Dehghannasiri, X. Qian, E.R. Dougherty, Intrinsically Bayesian robust Karhunen-Loève compression. *Signal Process.* **144**(Supplement C), 311–322 (2018)
21. T. Kailath, An innovations approach to least-squares estimation—part I: linear filtering in additive white noise. *IEEE Trans. Autom. Control.* **13**(6), 646–655 (1968)
22. A. Bryson, M. Frazier, in *Proceedings of the Optimum System Synthesis Conference*. Smoothing for linear and nonlinear dynamic systems (DTIC Document, Ohio, 1963), pp. 353–364
23. H. Rauch, Solutions to the linear smoothing problem. *IEEE Trans. Autom. Control.* **8**(4), 371–372 (1963)
24. H.E. Rauch, C. Striebel, F. Tung, Maximum likelihood estimates of linear dynamic systems. *AIAA J.* **3**(8), 1445–1450 (1965)
25. J.S. Meditch, Orthogonal projection and discrete optimal linear smoothing. *SIAM J. Control.* **5**(1), 74–89 (1967)
26. H. Zhao, P. Cui, W. Wang, D. Yang, h_∞ fixed-interval smoothing estimation for time-delay systems. *IEEE Trans. Signal Process.* **61**(2), 316–326 (2013)
27. B. Ait-El-Fquih, F. Desbouvries, On Bayesian fixed-interval smoothing algorithms. *IEEE Trans. Autom. Control.* **53**(10), 2437–2442 (2008)
28. D. Fraser, J. Potter, The optimum linear smoother as a combination of two optimum linear filters. *IEEE Trans. Autom. Control.* **14**(4), 387–390 (1969)
29. M. Briers, A. Doucet, S. Maskell, Smoothing algorithms for state-space models. *Ann. Inst. Stat. Math.* **62**(1), 61–89 (2010)
30. J.S. Meditch, On optimal linear smoothing theory. *J. Inf. Control.* **10**(6), 598–615 (1967)
31. T. Kailath, P. Frost, An innovations approach to least-squares estimation—part II: linear smoothing in additive white noise. *IEEE Trans. Autom. Control.* **13**(6), 655–660 (1968)
32. S. Nakamori, A. Hermoso-Carazo, J. Linares-Pérez, Design of a fixed-interval smoother using covariance information based on the innovations approach in linear discrete-time stochastic systems. *Appl. Math. Model.* **30**(5), 406–417 (2006)
33. S. Nakamori, A. Hermoso-Carazo, J. Linares-Pérez, M. Sánchez-Rodríguez, Fixed-interval smoothing problem from uncertain observations with correlated signal and noise. *Appl. Math. Comput.* **154**(1), 239–255 (2004)
34. F.R. Kschischang, B. J. Frey, A. H-Loeliger, Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory.* **47**(2), 498–519 (2001)
35. R.E. Griffin, A. P. Sage, Sensitivity analysis of discrete filtering and smoothing algorithms. *AIAA J.* **7**(10), 1890–1897 (1969)
36. S. Challa, M.R. Morelande, D. Musicki, R.J. Evans, *Fundamentals of object tracking*. (Cambridge University Press, Cambridge, 2011)
37. J.L. Williams, Marginal multi-Bernoulli filters: RFS derivation of MHT, JIPDA, and association-based MeMBeR. *IEEE Trans. Aerosp. Electron. Syst.* **51**(3), 1664–1687 (2015)
38. H. Liu, S. Zhou, H. Liu, H. Wang, in *Radar Conference (Radar), 2014 International*. Radar detection during tracking with constant track false alarm rate (IEEE, 2014), pp. 1–5
39. L. Qian, H. Wang, E.R. Dougherty, Inference of noisy nonlinear differential equation models for gene regulatory networks using genetic programming and Kalman filtering. *IEEE Trans. Signal Process.* **56**(7), 3327–3339 (2008)
40. M.S. Esfahani, E.R. Dougherty, Incorporation of biological pathway knowledge in the construction of priors for optimal Bayesian classification. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **11**(1), 202–218 (2014)

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com