Noise Covariance Identification for Nonlinear Systems using Expectation Maximization and Moving Horizon Estimation

Ming Ge [a], Eric C. Kerrigan [b],

[a] *Department of Electrical & Electronic Engineering, Imperial College London, London SW7 2AZ, UK*

[b] *Department of Electrical & Electronic Engineering and Department of Aeronautics, Imperial College London, London SW7 2AZ, UK*

Abstract

In order to estimate states from a noise-driven state space system, the state estimator requires a priori knowledge of both process and output noise covariances. Unfortunately, noise statistics are usually unknown and have to be determined from output measurements. Current expectation maximization (EM) based algorithms for estimating noise covariances for nonlinear systems assume the number of additive process and output noise signals are the same as the number of states and outputs, respectively. However, in some applications, the number of additive process noises could be less than the number of states. In this paper, a more general nonlinear system is considered by allowing the number of process and output noises to be smaller or equal to the number of states and outputs, respectively. In order to estimate noise covariances, a semi-definite programming solver is applied, since an analytical solution is no longer easy to obtain. The expectation step in current EM algorithms rely on state estimates from the extended Kalman filter (EKF) or smoother. However, the instability and divergence problems of the EKF could cause the EM algorithm to converge to a local optimum that is far away from true values. We use moving horizon estimation instead of the EKF/smoother so that the accuracy of the covariance estimation in nonlinear systems can be significantly improved.

Key words: Noise Covariance Estimation; Nonlinear System; Expectation Maximization; State Estimation; Full Information Estimation; Moving Horizon Estimation; Extended Kalman Filter

1 INTRODUCTION

For noise-driven state space systems, the performance of state estimation depends on properly defined noise statistics. Any guessed or inappropriately determined noise statistics may result in inaccuracies or even a divergence of state estimates. The pioneering work in [9] introduced the innovation correlation and the maximum likelihood estimation (MLE) methods for identifying the noise covariance.

The innovation correlation method is based on establishing an explicit relation between the noise covariance and the correlation of the innovation sequence. One of the algorithms for linear time invariant (LTI) systems is the auto-covariance least squares (ALS) method introduced in [13], in which the noise covariances are estimated by solving one semi-definite linear least squares problem. The ALS method was then extended to linear time-varying (LTV) as well as nonlinear systems in [5] and [12].

As an alternative approach, MLE aims to maximize the likelihood of the noise covariance given the output measurement sequence. An MLE-based "one-off" and an EM-based recursive algorithm for estimating noise covariances in LTI and LTV systems are presented in [22] and [17], respectively. The EM method was first introduced in [4] and was extended to nonlinear systems in [1] by using the extended Kalman filter (EKF) and Kalman smoother combined with the EM method.

In this paper, we consider more general nonlinear systems than [1] by allowing the number of process and output noises to be smaller or equal to the number of states and outputs, respectively. In our method, the noise covariance matrices are estimated using a semi-definite programming (SDP) solver, in order to ensure estimated noise covariances are positive definite. The

⋆ This paper was not presented at any IFAC meeting. Corresponding author Ming Ge. Tel. +44 (0) 7735198098.

Email addresses: ming.ge08@imperial.ac.uk (Ming Ge), e.kerrigan@imperial.ac.uk (Eric C. Kerrigan).

EM algorithm is combined with moving horizon estimation (MHE) and full information estimation (FIE) [16] to estimate noise covariances for nonlinear systems. The first reason for using MHE/FIE, rather than other Kalman-based filters, is to reduce the possibility of instability and divergence of estimated states. The second reason is that MHE/FIE allows one to add constraints, which could improve the accuracy of state estimation, if there exists physical limits or algebraic constraints on system states.

This paper is organized as follows: Sections 2 and 3 introduce the theory of covariance estimation using the EM algorithm combined with MHE/FIE for nonlinear systems. Two numerical examples are given in Section 4. We draw conclusions in Section 5.

$\mathbb{E}\{X|Y\}$ and $\mathbb{C}\{X|Y\}$ denote the conditional expected value and covariance of a random variable X given Y, respectively. $\mathcal{L}(\cdot)$ and $\mathcal{L}_\ell(\cdot)$ denote the likelihood and log-likelihood function of parameters of a statistical model, respectively. $(\cdot)^\dagger$ denotes the Moore-Penrose generalized inverse of a matrix. $\delta(x-m)$ is the Dirac delta function with delay m. $\mathscr{M}_l^{r,c}$ denotes an $r \times c$ auxiliary matrix that

$$\mathscr{M}_l^{r,c} := \left[\mathbf{0}_{r\times(l-1)}\ I_r\ \mathbf{0}_{r\times(c-r-l+1)}\right].$$

$|\cdot|$ and $\log|\cdot|$ denotes the determinant and log-determinant of a square matrix, respectively. $\mathscr{J}(\cdot)$ denotes the Jacobian matrix of a vector-valued function. $\|\cdot\|_F$ is the Frobenius norm of a matrix. $P \succ 0$ denotes that P is a positive-definite symmetric matrix; $\|x\|_W^2$ is weighted least squares of vector x, which equals to $x^\top W x$. $\mathrm{tr}(A)$ represents the trace of a square matrix A. The symbol $:=$ is to be read as 'is defined as', while $=:$ is to be read as 'defines'.

2 Noise Covariance Estimation using the EM Algorithm

Consider a discrete-time nonlinear state space model:

$$\begin{aligned} x_{k+1} &:= f(x_k) + G_k w_k \\ y_k &:= h(x_k) + H_k v_k \end{aligned} \tag{1}$$

where $x_k \in \mathbb{X}_k$ is the unknown state, $y_k \in \Re^p$ is the output measurement. $G_k \in \Re^{n\times r}$ and $H_k \in \Re^{p\times q}$ are two full column rank time-varying matrices in order to ensure the uniqueness of the conditional densities $p(x_{k+1}|x_k)$ and $p(y_k|x_k)$ to be defined later. w_k and v_k are two unknown noise terms, which affect the state and output, respectively.

Assumption 1 The noise sequences $(w_k)_{k=1}^M$ and $(v_k)_{k=1}^M$ are two random variables having Gaussian (or normal) distributions $\mathcal{N}(0,Q)$ and $\mathcal{N}(0,R)$, respectively, with zero mean and unknown positive-definite covariance matrices Q and R.

Assumption 2 The probability distribution $p(x_1)$ of the initial state has Gaussian distribution $p(x_1) \sim \mathcal{N}(\tilde{x}_1,\ P_1)$, where $\tilde{x}_1$ is the *a priori* most likely value of x_1 and P_1 is the corresponding error covariance.

Assumption 3 Functions $f(\cdot)$ and $h(\cdot)$ are twice differentiable. The discrete-time nonlinear model (1) is uniformly observable [10] and there exists a stable state observer for (1) with nonempty feasible region.

Define the full output and state sequence to be $Y_M := (y_k)_{k=1}^M$ and $X_M := (x_k)_{k=1}^M$, respectively. If Y_M and initial guesses of $\tilde{x}_1$, P_1, Q and R are all given, the set of true system parameters $\mathscr{O} := \{\tilde{x}_1, P_1, Q, R\}$ can be recursively estimated using the expectation maximization (EM) method [17].

Let the estimate of $\tilde{x}_1$ and covariance matrices P_1, Q and R at the i^{th} iteration be

$$\mathscr{O}_i := (\tilde{x}_{1,i}, P_{1,i}, Q_i, R_i), \quad i = 1,\ldots,N,$$

where $N \gg 1$ is the maximum number of iterations. The EM method recursively maximizes the expectation of the log-likelihood function $\mathscr{O}_i \mapsto \mathcal{L}_\ell(\mathscr{O}_i|Y_M)$, until the log-likelihood function converges to its maximum value [4].

The expression of $\mathscr{O} \mapsto \mathcal{L}_\ell(\mathscr{O}|Y_M)$ is given by

$$\begin{aligned} \mathcal{L}_\ell(\mathscr{O}|Y_M) &= \log\left(p(Y_M|\mathscr{O})\right) = \log\left(\frac{p(X_M,Y_M|\mathscr{O})}{p(X_M|Y_M,\mathscr{O})}\right) \\ &= \log\left(p(X_M,Y_M|\mathscr{O})\right) - \log\left(p(X_M|Y_M,\mathscr{O})\right). \end{aligned} \tag{2}$$

Taking the conditional expectation on both sides of (2) given Y_M and $\mathscr{O}_{i-1}$, we get the expectation of the log-likelihood function $\mathcal{L}_\ell(\mathscr{O}|Y_M)$, i.e.

$$\mathbb{E}\{\log(p(Y_M|\mathscr{O}))|Y_M,\mathscr{O}_{i-1}\} = \mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1}) - \mathcal{H}(\mathscr{O}|\mathscr{O}_{i-1}),$$

where $\mathcal{Q}$ and $\mathcal{H}$ are given by

$$\mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1}) := \mathbb{E}\{\log(p(X_M,Y_M|\mathscr{O}))|Y_M,\mathscr{O}_{i-1}\}, \tag{3a}$$
$$\mathcal{H}(\mathscr{O}|\mathscr{O}_{i-1}) := \mathbb{E}\{\log(p(X_M|Y_M,\mathscr{O}))|Y_M,\mathscr{O}_{i-1}\}. \tag{3b}$$

Because Y_M is a given measurement sequence,

$$\mathbb{E}\{\log(p(Y_M|\mathscr{O}))|Y_M,\mathscr{O}_{i-1}\} = \mathcal{L}_\ell(\mathscr{O}|Y_M).$$

Theorem 4 *[4,21] For the EM algorithm, the value of $\mathcal{L}_\ell(\mathscr{O}_i|Y_M)$ will monotonically increase at each iteration and converge to the maximum if*

$$\mathcal{Q}(\mathscr{O}_i|\mathscr{O}_{i-1}) \geq \mathcal{Q}(\mathscr{O}_{i-1}|\mathscr{O}_{i-1}), \quad \forall i, \tag{4}$$

with equality if and only if $\mathscr{O}_i = \mathscr{O}_{i-1}$.

Theorem 4 simplifies calculation, so that we only need to focus on fulfilling (4) by maximizing $\mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1})$, instead of $\mathcal{L}_\ell(\cdot|Y_M)$. Before giving the expression of $\mathcal{Q}$, we start with the expression of $\log(p(X_M, Y_M|\mathscr{O}))$ in (3a).

Lemma 5 *Consider two random vectors w, x and let $x = m + Gw$, where $m \in \Re^{n,1}$, $G \in \Re^{n,r}$ and $n > r$. If G is a full column rank matrix, $w \sim \mathcal{N}(\mathbf{0}, Q)$ and $Q \succ 0$, then the probability distribution of x is given by*

$$p(x) \propto \frac{1}{\sqrt{(2\pi)^r |W|}} e^{-\frac{1}{2}\|S(x-m)\|^2_{W^{-1}}},$$

where $W \succ 0$ and S is a constant matrix.

PROOF. Because G is a full column rank matrix, random vector x has a singular joint normal distribution, such that $p(x) \sim \mathcal{N}\left(\mathbf{0}, GQG^\top\right)$, where $GQG^\top \succeq 0$. By using the singular value decomposition, we have

$$GQG^\top = UPU^\top = U \begin{bmatrix} \tilde{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U^\top,$$

where $\tilde{Q} \in \Re^r$ is a nonsingular matrix, U is a unitary matrix, such that $U^{-1} = U^\top$. Define a new random vector z, such that

$$z := \begin{bmatrix} z_1^\top & z_2^\top \end{bmatrix}^\top = U^{-1}x = U^{-1}Gw,$$

where $z_1 \in \Re^{r\times 1}$. Since $z \sim \mathcal{N}(U^{-1}m, P)$, by definition of the singular joint normal distribution [6, pp. 376-377],

$$p(z) = \frac{\delta(z_2 - \mathscr{M}_{r+1}^{n-r,n} U^{-1} m)}{\sqrt{(2\pi)^r |\tilde{Q}|}} e^{-\frac{1}{2}\|z_1 - \mathscr{M}_1^{r,n} U^{-1} m\|^2_{\tilde{Q}^{-1}}},$$

Because $\delta(z_2 - \mathscr{M}_{r+1}^{n-r,n} U^{-1} m)$ is the probability mass function of the degenerate variable z_2, we have

$$p(z) \propto \frac{1}{\sqrt{(2\pi)^r |\tilde{Q}|}} e^{-\frac{1}{2}\|z_1 - \mathscr{M}_1^{r,n} U^{-1} m\|^2_{\tilde{Q}^{-1}}}.$$

Using transformation of variables [19, p. 22] gives

$$p(x) = p(z)\left|\mathscr{J}\left(U^{-1}x\right)\right| \propto \frac{1}{\sqrt{(2\pi)^r |W|}} e^{-\frac{1}{2}\|S(x-m)\|^2_{W^{-1}}}, \tag{5}$$

where $W = \tilde{Q}$ and $S = \mathscr{M}_1^{r,n} U^{-1}$. ■

Proposition 6 *If Assumptions 1 and 2 hold, then $\log(p(X_M, Y_M|\mathscr{O}))$ can be written as a function of x_1, $(w_k)_{k=1}^M$, $(v_k)_{k=1}^M$ and $\mathscr{O}$, such that*

$$\begin{aligned} \log(p(X_M, Y_M|\mathscr{O})) = \mathcal{L}_\ell\left(\mathscr{O}|X_M, Y_M\right) = \\ -\frac{1}{2}\left[s\log(2\pi) + \log|P_1| + M\log|Q| + M\log|R|\right] \\ -\frac{1}{2}\left[\varepsilon_1^\top P_1^{-1}\varepsilon_1 + \sum_{k=1}^{M-1} w_k^\top Q^{-1} w_k + \sum_{k=1}^{M} v_k^\top R^{-1} v_k\right]. \end{aligned} \tag{6}$$

where $s := n + r + p$ and $\varepsilon_k := x_k - \tilde{x}_k$.

PROOF. By using the relation between the joint and conditional probability stated earlier, we have

$$\log\left(p(X_M, Y_M|\mathscr{O})\right) \propto \log\left(p(X_M|Y_M, \mathscr{O})\right).$$

From Bayes' theorem we have

$$\log\left(p(X_M|Y_M, \mathscr{O})\right) \propto \log\left(p(Y_M|X_M, \mathscr{O}) p(X_M|\mathscr{O})\right),$$

where

$$\begin{aligned} p(Y_M|X_M, \mathscr{O}) &= \textstyle\prod_{k=1}^M p(y_k|x_k, \mathscr{O}), \\ p(X_M|\mathscr{O}) &= p(x_1) \textstyle\prod_{k=1}^{M-1} p(x_{k+1}|x_k, \mathscr{O}). \end{aligned}$$

By Assumption 1 and Lemma 5, we have

$$p(y_k|x_k, \mathscr{O}) \propto \frac{1}{\sqrt{(2\pi)^q |\tilde{R}_k|}} e^{-\frac{1}{2}\|\mathscr{M}_1^{q,p}\mathscr{Y}_k^{-1}(y_k - h(x_k))\|^2_{\tilde{R}_k^{-1}}},$$

$$p(x_{k+1}|x_k, \mathscr{O}) \propto \frac{1}{\sqrt{(2\pi)^r |\tilde{Q}_k|}} e^{-\frac{1}{2}\|\mathscr{M}_1^{r,n}\mathscr{X}_k^{-1}(x_{k+1} - f(x_k))\|^2_{\tilde{Q}_k^{-1}}},$$

where $\mathscr{X}_k$ and $\mathscr{Y}_k$ are unitary matrices, $\tilde{Q}_k,\ \tilde{R}_k \succ 0$,

$$\begin{aligned} G_k Q G_k^\top &= \mathscr{X}_k P_k^x \mathscr{X}_k^\top = \mathscr{X}_k \begin{bmatrix} \tilde{Q}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathscr{X}_k^\top, \\ H_k R H_k^\top &= \mathscr{Y}_k P_k^y \mathscr{Y}_k^\top = \mathscr{Y}_k \begin{bmatrix} \tilde{R}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathscr{Y}_k^\top. \end{aligned}$$

Rearranging $\|\mathscr{M}_1^{r,n}\mathscr{X}_k^{-1}(x_{k+1} - f(x_k))\|^2_{\tilde{Q}_k^{-1}}$ in (5) gives

$$\begin{aligned} &\|\mathscr{M}_1^{r,n}\mathscr{X}_k^{-1}(x_{k+1} - f(x_k))\|^2_{\tilde{Q}_k^{-1}} = \|\mathscr{M}_1^{r,n}\mathscr{X}_k^{-1} G_k w_k\|^2_{\tilde{Q}_k^{-1}} \\ &= w_k^\top G_k^\top \mathscr{X}_k^{-\top} (P_k^x)^\dagger \mathscr{X}_k^{-1} G_k w_k \\ &= w_k^\top G_k^\top \left(\mathscr{X}_k P_k^x \mathscr{X}_k^\top\right)^\dagger G_k w_k = w_k^\top G_k^\top \left(G_k Q G_k^\top\right)^\dagger G_k w_k \\ &= w_k^\top Q^{-1} w_k. \end{aligned} \tag{7a}$$

Similarly for the term $\|\mathscr{M}_1^{q,p}\mathscr{Y}_k^{-1}(y_k - h(x_k))\|^2_{\tilde{R}_k^{-1}}$ in (5),

$$\|\mathscr{M}_1^{q,p}\mathscr{Y}_k^{-1}(y_k - h(x_k))\|^2_{\tilde{R}_k^{-1}} = v_k^\top R^{-1} v_k. \tag{7b}$$

Substituting (7) into (5) and replacing $|\tilde{Q}_k|$ and $|\tilde{R}_k|$ with a linear function of $|Q|$ and $|R|$, respectively, gives

$$p(Y_M|X_M,\mathscr{O}) \propto \prod_{k=1}^{M} \frac{1}{\sqrt{(2\pi)^q|R|}} e^{-\frac{1}{2}v_k^\top R^{-1} v_k}$$
$$p(X_M|\mathscr{O}) \propto p(x_1) \prod_{k=1}^{M-1} \frac{1}{\sqrt{(2\pi)^r|Q|}} e^{-\frac{1}{2}w_k^\top Q^{-1} w_k}.$$

By taking the logarithm on both sides of $p(X_M|Y_M,\mathscr{O})$, we get (6). ■

Proposition 7 *For the nonlinear state space model* (1)*, if Assumption 3 holds, then* $\mathcal{Q}$ *defined in* (3a) *can be approximated by a function of* $\mathscr{O}$*,* $A_{k|M}$*,* $C_{k|M}$*,* $\hat{x}_{k|M}$*,* $P_{k|M}$ *and* $P_{k+1|k}$*, where*

$$A_{k|M} := \left.\frac{\partial f(\cdot)}{\partial x_k}\right|_{\substack{x_k=\hat{x}_{k|M}\\ w_k=\hat{w}_{k|M}}}, \quad C_{k|M} := \left.\frac{\partial h(\cdot)}{\partial x_k}\right|_{\substack{x_k=\hat{x}_{k|M}\\ v_k=\hat{v}_{k|M}}},$$
$$P_{a|b} := \mathbb{E}\left\{\left(x_a-\hat{x}_{a|b}\right)\left(x_a-\hat{x}_{a|b}\right)^\top \middle| Y_b, \mathscr{O}_{i-1}\right\}. \tag{8}$$

PROOF. The expression of $\mathcal{Q}$ for an LTV model was provided in [17]. We extend this method to the nonlinear model (1) via linearization. Taking the conditional expectation over (6) gives

$$\begin{aligned}\mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1}) &= \mathbb{E}\{\log(p(X_M,Y_M|\mathscr{O}))|Y_M,\mathscr{O}_{i-1}\} = \\ &-\frac{1}{2}\left[s\log(2\pi)+\log|P_1|+M\log|Q|+M\log|R|\right] \\ &-\frac{1}{2}\mathrm{tr}\Big\{P_1^{-1}\mathbb{E}\left[\varepsilon_1\varepsilon_1^\top|Y_M,\mathscr{O}_{i-1}\right]+Q^{-1}\times \\ &\sum_{k=1}^{M-1}\mathbb{E}\left[w_kw_k^\top|Y_M,\mathscr{O}_{i-1}\right]-R^{-1}\sum_{k=1}^{M}\mathbb{E}\left[v_kv_k^\top|Y_M,\mathscr{O}_{i-1}\right]\Big\},\end{aligned} \tag{9}$$

where

$$\begin{aligned}\mathbb{E}\left[w_kw_k^\top|Y_M,\mathscr{O}_{i-1}\right] &= \mathbb{C}\left[w_k|Y_M,\mathscr{O}_{i-1}\right] \\ &\quad+\mathbb{E}\left[w_k|Y_M,\mathscr{O}_{i-1}\right]\mathbb{E}\left[w_k|Y_M,\mathscr{O}_{i-1}\right]^\top, \\ &= \mathbb{E}\left[\left(w_k-\hat{w}_{k|M}\right)\left(w_k-\hat{w}_{k|M}\right)^\top\right]+\hat{w}_{k|M}\hat{w}_{k|M}^\top,\end{aligned} \tag{10a}$$

$$\begin{aligned}\mathbb{E}\left[v_kv_k^\top|Y_M,\mathscr{O}_{i-1}\right] &= \mathbb{C}\left[v_k|Y_M,\mathscr{O}_{i-1}\right] \\ &\quad+\mathbb{E}\left[v_k|Y_M,\mathscr{O}_{i-1}\right]\mathbb{E}\left[v_k|Y_M,\mathscr{O}_{i-1}\right]^\top, \\ &= \mathbb{E}\left[\left(v_k-\hat{v}_{k|M}\right)\left(v_k-\hat{v}_{k|M}\right)^\top\right]+\hat{v}_{k|M}\hat{v}_{k|M}^\top.\end{aligned} \tag{10b}$$

Hence, $\left(w_k-\hat{w}_{k|M}\right)$ and $\left(v_k-\hat{v}_{k|M}\right)$ can be approximated by applying first order Taylor series on (1), given Y_M and $\mathscr{O}_{i-1}$ [18, p. 408], such that

$$\begin{aligned}\left(v_k-\hat{v}_{k|M}\right) &\approx -H_k^\dagger C_{k|M}\left(x_k-\hat{x}_{k|M}\right), \\ \left(w_k-\hat{w}_{k|M}\right) &\\ &\approx G_k^\dagger\left[\left(x_{k+1}-\hat{x}_{k+1|M}\right)-A_{k|M}\left(x_k-\hat{x}_{k|M}\right)\right].\end{aligned}$$

Hence we have,

$$\mathbb{C}\left[v_k|Y_M,\mathscr{O}_{i-1}\right] = \mathcal{C}_k^R = H_k^\dagger C_{k|M}P_{k|M}C_{k|M}^\top H_k^{\dagger\top},$$
$$\mathbb{C}\left[w_k|Y_M,\mathscr{O}_{i-1}\right] = \mathcal{C}_k^Q = \tag{11}$$
$$\begin{aligned}&= G_k^\dagger\Big[\left(I_n-A_{k|M}U_k\right)P_{k+1|M}\left(I_n-A_{k|M}U_k\right)^\top \\ &\quad+A_{k|M}\left(P_k-U_kP_{k+1|k}U_k^\top\right)A_{k|M}^\top\Big]G_k^{\dagger\top},\end{aligned} \tag{12}$$

where $U_k := P_{k|k}A_k^\top P_{k+1|k}^{-1}$, $A_k := \frac{\partial f(\cdot)}{\partial x_k}\big|_{\substack{x_k=\hat{x}_k\\ w_k=0}}$ and

$$P_{k+1,k|M} = P_{k+1|M}U_{k-1}^\top = \mathbb{E}\left\{\left(x_{k+1}-\hat{x}_{k+1|M}\right)\left(x_k-\hat{x}_{k|M}\right)^\top\middle|Y_M,\mathscr{O}_{i-1}\right\}.$$

Substituting (10) and (11) into (9) gives

$$\begin{aligned}\mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1}) &= \mathbb{E}\{\log(p(X_M,Y_M|\mathscr{O}))|Y_M,\mathscr{O}_{i-1}\} \approx \\ &-\frac{1}{2}\left[s\log(2\pi)+\log|P_1|+M\log|Q|+M\log|R|\right] \\ &-\frac{1}{2}\mathrm{tr}\left\{P_1^{-1}\left[P_{1|M}+\left(\tilde{x}_1-\hat{x}_{1|M}\right)\left(\tilde{x}_1-\hat{x}_{1|M}\right)^\top\right]\right\} \\ &-\frac{1}{2}\mathrm{tr}\left(Q^{-1}\mathcal{C}_Q\right)-\frac{1}{2}\mathrm{tr}\left(R^{-1}\mathcal{C}_R\right).\end{aligned} \tag{13}$$

where $\mathcal{C}_Q := \sum_{k=1}^{M-1}\hat{w}_{k|M}\hat{w}_{k|M}^\top + \mathcal{C}_k^Q$ and $\mathcal{C}_R := \sum_{k=1}^{M-1}\hat{v}_{k|M}\hat{v}_{k|M}^\top + \mathcal{C}_k^R$. ■

Because $\tilde{x}_{1,i}$ and $P_{1,i}$ cannot be identified simultaneously, we let $\tilde{x}_{1,i} = \hat{x}_{1|M}$ and $P_{1,i} = P_{1|M}$ [17]. Thus, if the sequences $\left(\hat{x}_{k|M}\right)_{k=1}^{M}$, $\left(P_{k+1|k}\right)_{k=1}^{M-1}$ and $\left(P_{k|M}\right)_{k=1}^{M}$ are all computed using the EKF and the Kalman smoother algorithms (given in Section 3), then the set $\mathscr{O}_i$ can be estimated by iteratively solving

$$\begin{aligned}\mathscr{O}_i = \arg\max_{\substack{\tilde{x}_1,P_1,\\ Q,R}} \ &\mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1}) \\ \text{s.t. } &(P_1,\ Q,\ R)\succ 0.\end{aligned} \tag{14}$$

until both $\|Q_{i+1}-Q_i\|_F$ and $\|R_{i+1}-R_i\|_F$ are smaller than a positive threshold $\zeta \ll 1$. The problem (14) can be solved either by using a SDP solver, such as SDPT-3 [20], or by taking the partial derivative of (13) with respect to each entries of Q^{-1} and R^{-1}, and set it to zero, Q_i and R_i can be determined by [1]

$$Q_i = \mathcal{C}_Q/(M-1) \qquad \text{and} \qquad R_i = \mathcal{C}_R/M. \tag{15}$$

3 State Estimation using MHE and FIE

From the previous section, it follows that (6) is actually the objective function of an FIE problem [16] given Y_M and $\mathscr{O}_{i-1}$. Thus, it is natural to estimate the state and

noise sequence $\left(\hat{x}_{k|M}, \hat{v}_{k|M}, \hat{w}_{k|M}\right)_{k=1}^{M}$ by recursively solving the following problem, for $\bar{k} = 2, \cdots, M$:

$$X^*_{1,\bar{k}} := \arg\min_{X_{1,\bar{k}}} \|\varepsilon_1\|^2_{P^{-1}_{1,i-1}} + \sum_{k=1}^{\bar{k}-1} \|w_k\|^2_{Q^{-1}_{i-1}} + \sum_{k=1}^{\bar{k}} \|v_k\|^2_{R^{-1}_{i-1}} \tag{16a}$$

$$\text{s.t.} \quad x_{k+1} = f(x_k) + G_k w_k, \ k = 1, \ldots, \bar{k}-1 \tag{16b}$$

$$y_k = h(x_k) + H_k v_k, \quad k = 1, \ldots, \bar{k} \tag{16c}$$

$$x_k \in \mathbb{X}_k, \ k = 1, \ldots, \bar{k} \tag{16d}$$

where $X_{1,\bar{k}} := \begin{bmatrix} x_1^\top & v_1^\top & w_1^\top & x_2^\top & \cdots & w_{\bar{k}-1}^\top & x_{\bar{k}}^\top & v_{\bar{k}}^\top \end{bmatrix}^\top$ and the solution

$$X^*_{1,\bar{k}} =: \begin{bmatrix} \hat{x}^\top_{1|\bar{k}} & \hat{v}^\top_{1|\bar{k}} & \hat{w}^\top_{1|\bar{k}} & \cdots & \hat{w}^\top_{\bar{k}-1|\bar{k}} & \hat{x}^\top_{\bar{k}} & \hat{v}^\top_{\bar{k}} \end{bmatrix}^\top,$$

where $(\hat{x}_{k|\bar{k}})_{k=1}^{\bar{k}}$, $(\hat{w}_{k|\bar{k}})_{k=1}^{\bar{k}-1}$ and $(\hat{v}_{k|\bar{k}})_{k=1}^{\bar{k}}$ are the estimates of the state and system noise sequence $(x_k)_{k=1}^{\bar{k}}$, $(w_k)_{k=1}^{\bar{k}-1}$ and $(v_k)_{k=1}^{\bar{k}}$, respectively, with $\hat{(\cdot)}_k := \hat{(\cdot)}_{k|k}$, given output sequence $Y_{\bar{k}}$.

For the FIE, the problem size grows with time as the estimator processes more data; as a result, the problem complexity scales at least linearly with M. In order to make the estimation problem tractable, we need to band the problem size by using MHE [16], such that

$$\Theta^*_{\bar{k},i} := \min_{X_{k_s,\bar{k}}} \|\varepsilon_{k_s}\|^2_{P^{-1}_{k_s}} + \sum_{k=k_s}^{\bar{k}-1} \|w_k\|^2_{Q^{-1}_i} + \sum_{k=k_s+1}^{\bar{k}} \|v_k\|^2_{R^{-1}_i} \tag{17a}$$

$$\text{s.t.} \quad x_{k+1} = f(x_k) + G_k w_k, \ k = k_s, \ldots, \bar{k}-1 \tag{17b}$$

$$y_k = h(x_k) + H_k v_k, \ k = k_s, \ldots, \bar{k} \tag{17c}$$

$$x_k \in \mathbb{X}_k, \ k = k_s, \ldots, \bar{k} \tag{17d}$$

where $k_s := \max\{\bar{k} - N_l, 0\} + 1$; N_l is a positive integer called the horizon length and $N_l \leq M$. The MHE is equivalent to FIE if $\bar{k} = N_l$; for $\bar{k} \leq N_l$, $\hat{x}_1 := \tilde{x}_1$ and $P_{k_s} := P_{1,i}$. $\varepsilon^\top_{k_s} P^{-1}_{k_s} \varepsilon_{k_s}$ is called arrival cost.

A stable algorithm for determining the arrival cost without any linearization or approximation is given in [14, p. 252], which establishes the accompanying MHE problem [14, p. 252]:

$$\Theta^{0*}_{\bar{k},i} := \min_{X^0_{k_s,\bar{k}}} + \sum_{k=k_s}^{\bar{k}-1} \|w_k\|^2_{Q^{-1}_i} + \sum_{k=k_s+1}^{\bar{k}} \|v_k\|^2_{R^{-1}_i}$$

$$\text{s.t.} \quad \text{(17b)}, \ \text{(17c)}, \ \text{(17d)}$$

where $X^0_{k_s,\bar{k}} := X_{k_s,\bar{k}}$ and the solution

$$X^{0*}_{k_s,\bar{k}} =: \begin{bmatrix} \hat{x}^{0\top}_{k_s|\bar{k}} & \hat{v}^{0\top}_{k_s|\bar{k}} & \hat{w}^{0\top}_{k_s|\bar{k}} & \cdots & \hat{w}^{0\top}_{\bar{k}-1|\bar{k}} & \hat{x}^{0\top}_{\bar{k}} & \hat{v}^{0\top}_{\bar{k}} \end{bmatrix}^\top.$$

The weight P_{k_s} in the arrival cost is defined by $P_{k_s} := \gamma_{k_s} P_{1,i}$, where $\gamma_{k_s} \in [0,\ 1]$ and is determined by finding the maximum possible γ_{k_s} that satisfies the stability condition

$$\gamma^{-1}_{k_s} \|\varepsilon_{k_s}\|^2_{P^{-1}_{1,i}} + \Theta^*_{\bar{k}-N_l,i} \leq \sum_{k=0}^{m} \Theta^{0*}_{k_s-kN_l,i} + \|\varepsilon_1\|^2_{P^{-1}_{1,i}},$$

where m is the quotient of k_s divided by N_l.

Note that, for the MHE algorithm stated above, if Assumption 3 holds, $\mathbb{X}_1$ is a compact set and the disturbances w_k and v_k are all bounded, then the state estimation errors $\|x_k - \hat{x}_k\|_{k=1}^{M}$ are guaranteed to be bounded for all $k > N_l$ [14, Prop. 3.8].

Proposition 8 *For the nonlinear system* (1)*, given $Y_{\bar{k}}$ and noise covariances, if Assumption 2 holds, then the state estimation using the FIE* (16) *will have a smaller estimation error $e_{\bar{k}} := \|x_{\bar{k}} - \hat{x}_{\bar{k}}\|$ compared to the EKF.*

PROOF. In [8], it was proved that the EKF is equivalent to solving the FIE problem (16) by only one Newton step with a carefully chosen initial guess. Hence, by optimality, the state estimate using the FIE will have a smaller error $e_{\bar{k}}$ than the EKF. ■

After $\left(\hat{x}_{k|M}\right)_{k=1}^{M}$ is obtained from (16), $\left(P_{k|M}\right)_{k=1}^{M}$ and $\left(P_{k+1|k}\right)_{k=1}^{M-1}$ in (13) can both be determined by the EKF [8] and Kalman smoother [17] equations. We formulate the noise covariance estimation strategy in Algorithm 1, which will iterate until the condition at step 8 is satisfied. The main problem with Algorithm 1 is that step 3 could be time-consuming, especially when the number of decision variables and the horizon length are large. Instead of using an MHE, an EKF/EKS could be used to provide the initial guess of (16) with

$$z_0 := \begin{bmatrix} \hat{x}^\top_{1|M} & \mathbf{0} & \hat{x}^\top_{2|M} & \cdots & \hat{x}^\top_{M} & \mathbf{0} & \hat{x}^\top_{M+1|M} \end{bmatrix}^\top,$$

with the assumption that $A_k \approx A_{k|M}$ and $C_k \approx C_{k|M}$, which we call "Fast Algorithm 1".

Note that, for LTV systems, [8] shows that the FIE and the Kalman filter are two equivalent algorithms and it is also straightforward to show the equivalence between the FIE and the Kalman smoother [15]. Therefore, for LTV systems, our algorithm will result in the EM methods for linear systems introduced in [17].

A different approach for estimating the covariances for nonlinear systems is provided in [1], which replaces steps 3 to 5 in Algorithm 1 by EKF/EKS, assumes $G_k = I_n$, $H_k = I_p$ and provides an equivalent expression ((49)–(55) in [1]) for $\mathcal{C}_Q$ and $\mathcal{C}_R$. For determining the noise covariances, an analytical expression (15) is used in [1].

Algorithm 1 Noise Covariance Estimation Algorithm: $\mathscr{O} = \text{EM-FIE}\ (\mathscr{O}_1,\ Y_M,\ N_l,\ \zeta)$

Let $b \leftarrow 0$ and $i \leftarrow 1$,
while $b = 0$ **do**
  Recursively estimate $(\hat{w}_k)_{k=1}^{M-1}$ and $(\hat{x}_k, \hat{v}_k)_{k=1}^{M}$ using the MHE with horizon length N_l
  Let $z_0 := \begin{bmatrix} \hat{x}_1^\top & \hat{v}_1^\top & \hat{w}_1^\top & \hat{x}_2^\top & \cdots & \hat{w}_{M-1}^\top & \hat{x}_M^\top & \hat{v}_M^\top \end{bmatrix}^\top$
  Obtain $\left(\hat{x}_{k|M}\right)_{k=1}^{M}$ by solving (16) using z_0 as the initial guess
  Calculate $\left(P_{k|M}\right)_{k=1}^{M}$ and $\left(P_{k+1|k}\right)_{k=1}^{M-1}$ using the Kalman smoother equations
  Obtain $\mathscr{O}_{i+1}$ by solving (14)
  if $\|Q_{i+1} - Q_i\|_F < \zeta$ and $\|R_{i+1} - R_i\|_F < \zeta$ **then**
    Let $b \leftarrow 1$
  else
    $i \leftarrow i + 1$, $P_{1,i+1} \leftarrow P_{1|M}$ and $\tilde{x}_{1,i+1} \leftarrow \hat{x}_{1|M}$
  end if
end while

The main problem is that (49)–(55) in [1] involve more addition of terms with opposite sign (plus/minus) compared to (13), which may affect the estimation accuracy due to numerical errors, hence could cause non-positive definite Q_i and R_i, if the analytical expression (15) is used. In order to improve the estimation accuracy and guarantee the positive definiteness of estimated noise covariances, one should replace (49)–(55) in [1] and (15) with (13) and an SDP solver, respectively. More details about this problem will be discussed in Section 4.2.

The main advantage of using the EKF/EKS is that this is faster than the MHE/FIE; however, the EKF/EKS sometimes suffers from instability and accuracy problems. By using the MHE algorithm given in [14, p. 252], stability can be guaranteed by [14, Prop. 3.8]. Proposition 8 will ensure the MHE/FIE can always provide more accurate state estimates $(\hat{x}_k)_{k=1}^{M}$ than the EKF. Moreover, if there exists state constraints in the original system, then adding constraints in the MHE/FIE will significantly improve the estimation accuracy.

4 Numerical Example

4.1 Nonlinear Example 1

Consider a nonlinear system:

$$\underbrace{\begin{bmatrix} a_{k+1} \\ b_{k+1} \\ c_{k+1} \end{bmatrix}}_{x_{k+1}} := \underbrace{\begin{bmatrix} +a_k \cos(c_k T_s) + b_k \sin(c_k T_s) \\ -a_k \sin(c_k T_s) + b_k \cos(c_k T_s) \\ c_k \end{bmatrix}}_{f(x_k)} + G \underbrace{\begin{bmatrix} w_k^a \\ w_k^b \end{bmatrix}}_{w_k}$$
$$y_k := Cx_k + v_k \tag{18}$$

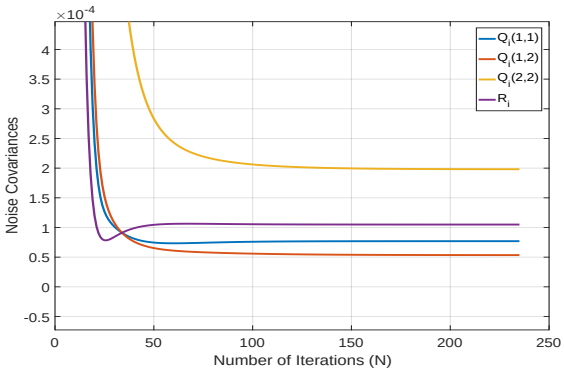


Fig. 1. Estimates of the process noise covariance

where $T_s = 0.1\,\text{sec}$ is the sampling time and

$$C := \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad G := \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}^\top.$$

$w_k \in \Re^{2\times 1}$ and $v_k \in \Re^1$ satisfy Assumption 1, which w_k^a and w_k^b are uncorrelated. We now pick an initial state $x_1 := \begin{bmatrix} 3 & -2 & -0.5 \end{bmatrix}^\top$, then generate output measurements $(y_k)_{k=1}^{1000}$ based on noise covariance matrices

$$Q = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \times 10^{-4},\ R = 1 \times 10^{-4}.$$

The guessed state error covariance is $P_{1,1} = 0.1 \times I_3$ and the guessed initial state is $\hat{x}_{1|0} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^\top$, the guessed noise covariance matrices are set to $Q_g = I_2$ and $R_g = 1$. Figure 1 is the covariance estimation results for nonlinear system (18) using the EM/EKS formulation (13) and SDP solver. Figure 1 shows that after several iteration, the estimated noise covariances converge to

$$\hat{Q} = \begin{bmatrix} 0.77 & 0.53 \\ 0.53 & 2 \end{bmatrix} \times 10^{-4},\ \hat{R} = 1 \times 10^{-4},$$

respectively. Figure 2 is a plot of the log-likelihood function $\mathcal{Q}$ versus the number of iterations.

4.2 Nonlinear Example 2

Consider the nonlinear chemical process [7]:

$$\underbrace{\begin{bmatrix} x_{k+1}^{(1)} \\ x_{k+1}^{(2)} \end{bmatrix}}_{x_{k+1}} := \underbrace{\begin{bmatrix} \frac{x_k^{(1)}}{1+2k_r T_s x_k^{(1)}} \\ x_k^{(2)} + \frac{k_r T_s x_k^{(1)}}{1+2k_r T_s x_k^{(1)}} \end{bmatrix}}_{f(x_k)} + \underbrace{\begin{bmatrix} w_k^{(1)} \\ w_k^{(2)} \end{bmatrix}}_{w_k},$$
$$y_k := Cx_k + v_k \tag{19}$$

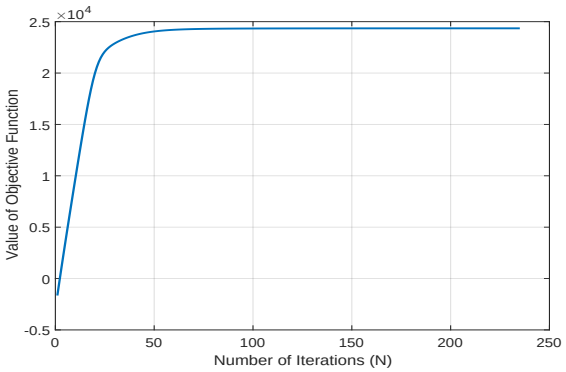


Fig. 2. Log-likelihood function $\mathcal{Q}$ versus number of iterations

where $x_k \in \Re^{2\times 1}$, $k_r = 0.16$, the sampling period $T_s = 0.01\,\text{s}$, $C = [1 \;\; 1]$, the initial state $x_0 = [3 \;\; 1]^\top$, $w_k \in \Re^{2\times 1}$ and $v_k \in \Re^1$ satisfy Assumption 1, which $w_k^{(1)}$ and $w_k^{(2)}$ are uncorrelated. We generate 7.5 seconds of output measurements $(y_k)_{k=1}^{750}$ based on noise covariance matrices $Q = 10^{-6} \times I_2$ and $R = 1 \times 10^{-2}$.

We will use Algorithm 1, Fast Algorithm 1 and the EKF/EKS based EM method provided in [1] to estimate the true noise covariances Q and R. Because noise covariances are determined by an SDP solver rather than an analytical expression (15), we are able to add an extra constraint such that decision variable Q in (14) has zero off-diagonal entries. For easier comparison, rather than setting up a terminal condition ζ, the maximum number of iterations N is fixed to 3000. Let $N_l = 200$, the initial guess of $P_{1,1}$, Q_1 and R_1 are $50\times I_2$, $Q_1 = I_2$ and $R_1 = 1$, respectively, with initial state guess $\tilde{x}_1 = [1 \;\; 45]^\top$.

Figure 3 is the plot of the smoothed state estimates using FIE and EKS with guessed noise covariance. Figure 3 clearly shows that FIE converges to the neighborhood of the true state trajectory very quickly, while EKS estimates are still far away from the true state trajectory. The main reason behind this is because FIE is able to add the constraint on the system state x_k that the partial pressure x_k is non-negative.

Figures 4 to 6 are the plots of the system noise covariance estimates using Alg. 1, Fast Alg. 1 and the EKF/EKS based EM method, respectively, over 3000 iterations. For the EKF/EKS based EM method, both SDP solver and (15) provided in [1] are used to determine noise covariances. In Figures 4 and 6, the estimated noise covariances using Alg. 1 and Fast Alg. 1 are almost identical and much closer to the true values, compared with results using the EKF/EKS based EM method. The estimated Q_i and R_i using (15) in Figures 5 and 6 are much worse than others, this is because the positive definiteness of Q_i using (15) has been violated several times due to the numerical errors.

Figure 7 is a plot of the log-likelihood function $\mathcal{Q}$ versus

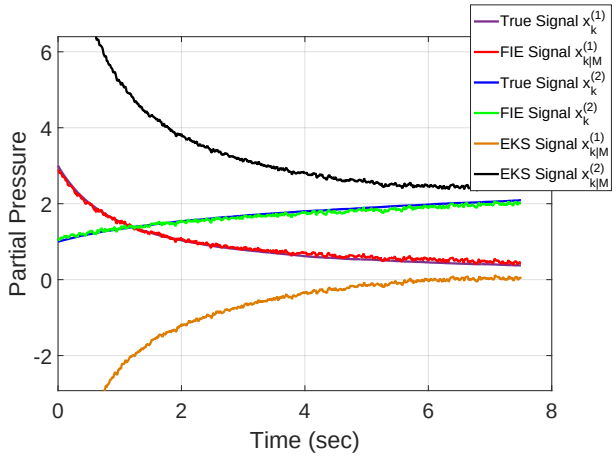


Fig. 3. Estimates of the state trajectory using Q_1 and R_1

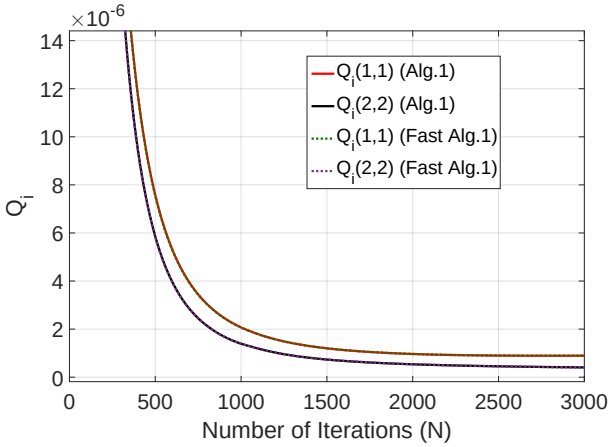


Fig. 4. Estimates of the process noise covariance

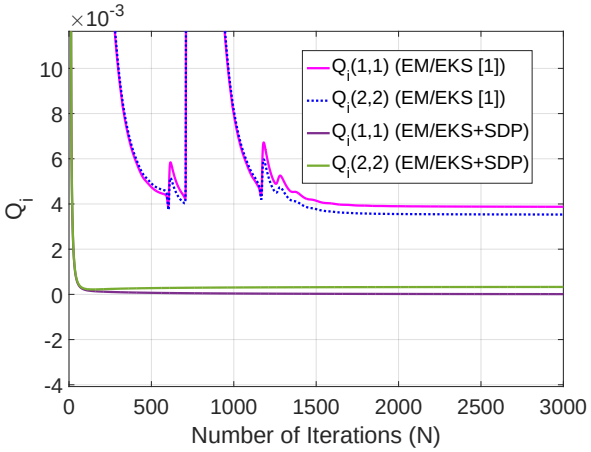


Fig. 5. Estimates of the process noise covariance using EKF/EKS in [1]

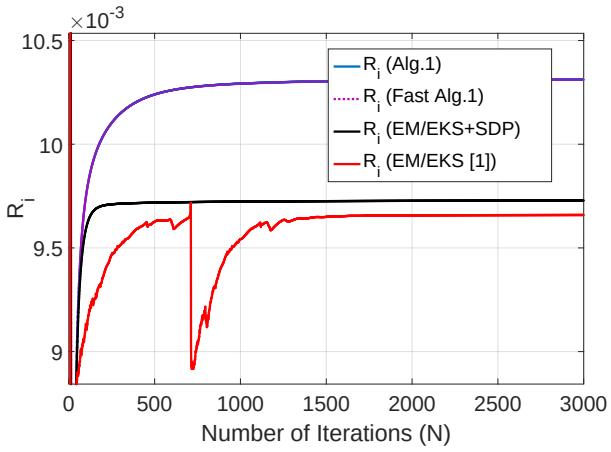


Fig. 6. Estimates of output noise covariance

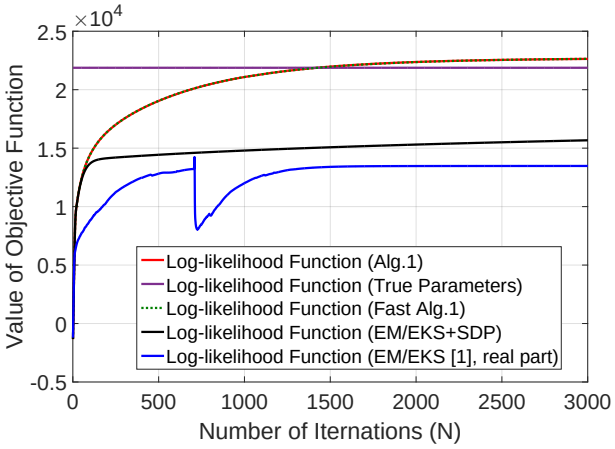


Fig. 7. Log-likelihood function $\mathcal{Q}$ versus number of iterations

the number of iterations. For the EKF/EKS method using (15), only the real part of complex $\mathcal{Q}$ is shown. Figure 7 shows that only the trajectories using Alg. 1 and fast Alg. 1 are asymptotically converging to the neighborhood of the reference line, which is calculated using the function $\mathcal{Q}$ with true covariances.

Since the noise covariance might be a square matrix rather than a scalar, we examine the estimation accuracy by calculating the "error norm" and the"error percentage", such that

$$e_Q := \|Q - Q^*\|_F, \quad e_R := \|R - R^*\|_F,$$
$$p_Q := \frac{e_Q}{\|Q\|_F} \times 100\ \%, \quad p_R := \frac{e_R}{\|R\|_F} \times 100\ \%,$$

where Q^* and R^* are the estimated process and output noise covariance matrices, respectively.

It is also interesting to compare the EM method with the ALS method introduced in [5]. We increase the data length M to 1000 and the sub-optimal filter gains $(L_k^s)_{k=1}^M$ and the state error covariance $(P_k)_{k=1}^M$ are obtained from the EKF equations from the same initial conditions $\hat{x}_{1|0} = \tilde{x}_1$, $P_{1|0} = P_{1,1}$ and guessed noise covariance matrices $Q_g = Q_1$ and $R_g = R_1$. We regenerate output measurements from the same initial state $x_0 = [3\ \ 1]^\top$ and different noise sequences based on noise covariances Q and R, hence repeat the simulation 200 times using both the EKF and MHE methods mentioned in [5] with $N = 50$, $M_e = 1000$ and $N_l = 250$, and calculate the average values of e_Q and e_R over 200 simulations as well as the error percentages.

Table 1
Noise covariance estimation results

	$e_Q(10^{-6})$	$e_R(10^{-4})$	$p_Q(\%)$	$p_R(\%)$
Alg. 1	0.603	1.88	42.43	1.88
Fast Alg. 1	0.607	1.88	42.67	1.88
EM/EKS SDP	348.5	2.713	24645	2.71
EM/EKS [1]	7396	3.412	1.2×10^6	3.8
	$\bar{e}_Q(10^{-3})$	$\bar{e}_R(10^{-4})$	$\bar{p}_Q(\%)$	$\bar{p}_R(\%)$
ALS/EKF	6.4	32	4×10^5	32.15

Table 1 compares the estimation error norms and percentages using different noise covariance estimation methods. For the nonlinear system (19), Table 1 clearly shows that noise covariance estimation using Alg. 1 and fast Alg. 1 give almost identical results, which are more accurate than the ALS based methods. The reason behind this is that the ALS method estimates noise covariances by minimizing the difference between the auto-covariance of the estimates (as a function of Q and R) and true innovation sequence. Unfortunately, if the system is nonlinear with a bad initial guess of noise covariances, then the auto-covariance of the estimated innovation sequence may not be approximated by a linear function of true noise covariances, which will significantly affect the ALS estimations. On the other hand, although our EM based algorithms involve approximations of the $\mathcal{Q}$ function and the global optimum is not guaranteed for any initial guess [21], the EM method will keep increasing the likelihood function $\mathcal{L}_\ell(\cdot|Y_M)$ until it has converged. This property could help us to find a better estimate of noise covariances for some nonlinear systems. Table 2 compares the computation time in minutes [1] using 5 different estimation methods shown in Table 1.

As we discussed at the end of Section 3, Table 3 examines the numerical robustness based on the EM+EKS method with different ways to maximize the $\mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1})$. T_{np} is number of times that Q_i are NOT positive definite within 3000 iterations. Table 3 shows that by replacing

[1] Based on an Intel i7 5960X Octa-core CPU at 4.43 GHz, MATLAB 2015a, IPOPT v3.12.1 and SDPT3 v4.0 solver.

Table 2
Computation Time (mins)

Alg. 1	Fast Alg. 1	EM/EKS SDP	EM/EKS [1]	ALS/EKF
3543	41.7	22.1	5.4	0.4

Table 3
Robustness tests for EM+EKS based noise covariance estimation

Method for maximizing $\mathcal{Q}(\mathscr{O}\|\mathscr{O}_{i-1})$	T_{np}	$p_Q(\%)$	$p_R(\%)$
EM/EKS [1]	11	1.2×10^6	3.8
EM/EKS (13)+(15)	0	73604	0.35

(49)–(55) in [1] with (13), all the Q_i and R_i within 3000 iterations are positive definite and has a smaller error percentage p_Q and p_R, respectively.

5 Conclusions

We have introduced a new EM-based noise covariance estimation algorithm for nonlinear systems, in which the state and output error terms can be written as an injective linear transformation of process and output noises, respectively; we also used the MHE/FIE to implement the expectation step in the EM method. The first nonlinear example was used to verify the EM/EKS formulation (13), when G is not square but full column rank matrix. The second nonlinear example was used to investigate the performance of the algorithm and we compared our EM method with the EM/EKS method introduced in [1] and the ALS algorithm introduced in [5]. Results showed that only our EM method is able to provide accurate noise covariance estimates. Compared with the algorithm introduced in [1], our method used the MHE and FIE instead of the EKF and EKS to estimate the system states, which took longer, but is able to obtain a much more accurate covariance estimate. We established an expression of $\mathcal{Q}(\mathscr{O}|\mathscr{O}_{i-1})$ that is numerically more robust — we also used a semi-definite programming (SDP) solver to estimate the noise covariances, in order to ensure the positive definiteness of the estimated covariances. A fast algorithm was introduced, which used EKS instead of MHE to provide an initial guess for the FIE. The example has shown that our fast algorithm will speed up the EM process by about 50 times compared to Alg. 1, with almost identical results.

Future work could involve improving the computational speed of the algorithm presented here, so that the method can be used in various fields for better state estimation and prediction results, such as chemical engineering and motion tracking. The most computationally intensive part is solving the SDP (14) and the resulting NLP (16) or (17) that arises in the FIE or MHE, respectively. It is well-known that, in general, the computational complexity of each iteration of a Newton-based interior point method scales at worst cubically with the size of the data of the optimization problem, and that the number of iterations for convergence to an optimum is roughly a constant [3,11]. This implies that the computational effort needed to solve the SDP (14) scales at worst to the sixth power of the number of states and noise terms. However, the optimization problems here are highly structured, hence the computational complexity of efficient tailor-made optimization solvers, such as those for dynamic estimation problems [2], can be made to scale significantly better than using general-purpose solvers. For example, if care is taken in using sparse linear algebra when computing the search directions, the complexity of computing a direct factorization of the linearized KKT system of (16) and (17) scales linearly in the horizon length and linearly to quadratically, at worst, in the number of states and noise/disturbance signals [23]. Note also that (14) is a convex optimization problem, hence all local solutions are globally optimal. However, (16) and (17) are non-convex, in general, so only convergence to a local minimum can be guaranteed with most gradient-based solvers. If global solutions are sought, instead, then global solvers can be used, but this will clearly result in a significant increase in computational effort.

References

[1] Vinay A. Bavdekar, Anjali P. Deshpande, and Sachin C. Patwardhan. Identification of process and measurement noise covariance for state and parameter estimation using extended Kalman filter. *Journal of Process Control*, 21(2011):585–601, 2011.

[2] J. Betts. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*. SIAM, 2nd edition, 2010.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[5] Ming Ge and Eric C. Kerrigan. Noise covariance estimation for time-varying and nonlinear systems. In *Proceeding of The* 19$^{\text{th}}$ *IFAC World Congress*, Cape Town, South Africa, August 2014.

[6] Michael D. Graham and James B. Rawlings. *Modelling and Analysis Principles for Chemical and Biological Engineers*. Nob Hill Publishing, 2013.

[7] Eric L. Haseltine and James B. Rawlings. A critical evaluation of extended Kalman filtering and moving horizon estimation. Technical Report 2002-03, University of Wisconsin-Madison, March 2003.

[8] Jeffrey Humpherys, Preston Redd, and Jeremy West. A fresh look at the Kalman filter. *SIAM Review*, 54(4):801–823, 2012.

[9] R. K. Mehra. Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, 15(12):903–908, 1972.

[10] P. E. Moraal and J. W. Grizzle. Observer design for nonlinear systems with discrete-time measurement. *IEEE Transactions on Automatic Control*, 40(3):395–404, 1995.

[11] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[12] M. R. Rajamani and J. B. Rawlings. Application of a new data-based covariance estimation technique to a nonlinear industrial blending drum. Technical Report 2007-03, University of Wisconsin-Madison, September 2007.

[13] M. R. Rajamani and J. B. Rawlings. Estimation of the disturbance structure from data using semidefinite programming and optimal weighting. *Automatica*, 45(1):142–148, 2009.

[14] Christopher. V. Rao, James. B. Rawlings, and David Q. Mayne. Constrained state estimation for nonlinear discrete-time system: Stability and moving horizon approximation. *IEEE Transactions on Automatic Control*, 48(2):246–258, 2003.

[15] H. E. Rauch, F. Tung, and C. T. Striebel. AIAA journal. *Maximum Likelihood Estimates of Linear Dynamic System*, 3(80):1445–1450, 1965.

[16] James B. Rawlings and David Q Mayne. *Model Predictive Control: Theory and Design*. Nob Hill Publishing, 2009.

[17] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

[18] Dan Simon. *Optimal State Estimation: Kalman, H_∞ and Nonlinear Approaches*. John Wiley & Son, Inc., New Jersey, U.S., 2006.

[19] Hisashi Tanizaki. *Computational Methods in Statistics and Econometrics*. Marcel Dekker, Inc., New York, U.S., 2004.

[20] K.C. Toh, M.J. Todd, and R.H. Tutuncu. SDPT3 — a MATLAB software package for semidefinite programming. *Optimization Methods and Software*, 11(1999):545–581, 1999.

[21] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[22] Megan A. Zagrobelny and James B. Rawlings. Identifying the uncertainty structure using maximum likelihood estimation. In *Proc. 2015 American Control Conference*, Chicago, U.S, July 2015.

[23] V. M. Zavala, C. D. Laird, and L. T. Biegler. A fast moving horizon estimation algorithm based on nonlinear programming sensitivity. *Journal of Process Control*, 18(9):876 – 884, 2008.