

Some quantitative characteristics of error covariance for Kalman filters

By WEI KANG^{1*}, and LIANG XU², ¹*Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA, USA;* ²*Naval Research Laboratory, Monterey, CA, USA*

(Manuscript Received 25 August 2020; in final form 13 November 2020)

ABSTRACT

Some *quantitative characteristics* of error covariance are studied for linear Kalman filters. These quantitative characteristics include the peak value and location in the matrix, the decay rate from peak to bottom, and some algebraic constraints of the elements in the covariance matrix. We mathematically prove a matrix upper bound and its quantitative characteristics for the error covariance of Kalman filters. Computational methods are developed to numerically estimate the elements in a matrix upper bound and its decay rate. The quantitative characteristics and the computational methods are illustrated using three examples, two linear systems and one nonlinear system of shallow water equations.

Keywords: error covariance, Kalman filter, quantitative characteristics, localisation

1. Introduction

A fundamental challenge in data assimilation lies in the computationally intractable process of estimating and updating the error covariance for high dimensional problems. Various methodologies, from ensemble-based methods to approaches of hybrid error covariance, have been extensively studied in the literature (Houtekamer and Zhang, 2016). Due to the limitation of computational power, the error covariance has to be estimated based upon a relatively small number of ensemble members. Rank deficiency, underestimation, and sampling errors are some notable drawbacks of these approaches. As a result, additional numerical treatments, such as covariance localisation and variance inflation, are required to improve the stability of algorithm and estimation accuracy. Some key parameters are critical for the effectiveness of these numerical treatments, for instance the radius of localisation and the magnitude of inflation. These parameters are estimated based on physical and statistical grounds and are validated through numerical experiments (Bannister, 2008a, 2008b; Ménétrier et al., 2015).

In Gaspari and Cohn (1999), the magnitude and decay of error covariances are quantitatively outlined using correlation functions. The result has been widely used for the localisation of ensemble Kalman filters (EnKFs).

Based on Schur product theorem, the element-wise multiplication of two positive definite matrices guarantees that the resulting symmetric matrix is a positive definite matrix itself. However, the underlying relationship between the parameter in a correlation function and the system model to be estimated is not clear. In this paper, we study some closely related subjects, the magnitude and decay of elements in error covariance, but taking a different approach. We derive and prove some *quantitative characteristics* of error covariance that are determined by various parts in a system model, including its dynamic model, the observation model and the covariances of random variables representing model uncertainties and observation error. The quantitative characteristics studied in this paper include the *peak value and its location* in the matrix, the *decay rate* from peak to bottom, the *approximate sparsity pattern*, etc.

The contributions of this paper consist of three parts. (i) We mathematically prove a theorem and several propositions on the quantitative characteristics of error covariance for linear Kalman filters. The results unveil some interconnections between system models and error covariance. (ii) We develop computational methods so that quantitative characteristics can be numerically computed. (iii) We demonstrate some potential applications of the discoveries through three examples. In the first example, we sketch the outline shape of error covariance using its peak and bottom value as well as the curve of a decay

*Corresponding author. e-mail: wkang@nps.edu

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

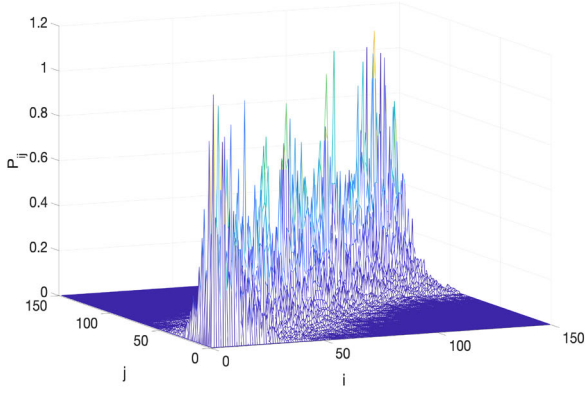
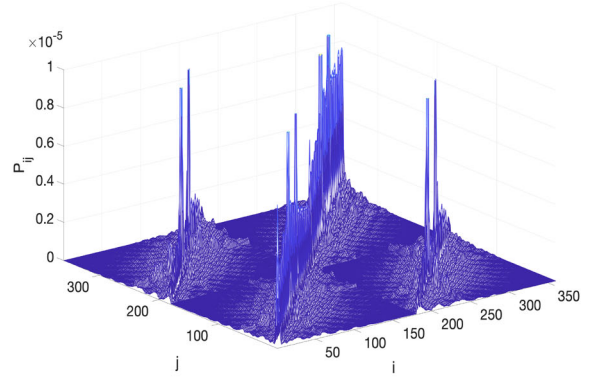


Fig. 1. Examples of Kalman filter error covariance.

function. The method's goal is not about replacing the need of estimating the error covariance in a process, such as EnKF or 4D-Var. Instead, the outline shape provides information that helps in the validation of an estimated error covariance if the full matrix is intractable or if its estimation is based on a relatively small ensemble size. In the second example, the quantitative characteristics of error covariance are used as an indicator of the upper bound for the localisation radius of an EnKF. In the third example, the results proved for linear systems are tested using a nonlinear model, the shallow water equations.

The theoretical approach in this paper is based on the use of the controllability Gramian of dynamical systems (see, for instance, Kailath (1980)). The concept has been widely used in control theory to measure the impact of a system's inputs, such as the model uncertainty in this study, on the system's trajectories. We prove that the error covariance of the Kalman-Bucy filter equals the difference between two controllability Gramians, in addition to the propagation of the initial background error covariance. This relationship leads to a matrix upper bound of the error covariance. Based upon the concept of duality in the control theory, each element in the upper bound matrix can be computed individually without the need of estimation of the entire error covariance matrix of a Kalman filter. This property makes the computational algorithms in this paper being *component-based*, i.e. the quantitative characteristics of error covariance can be computed over any given component in the matrix, such as a submatrix, a block of any shape, a row/column, etc. It is also proved that, if the observation measures a state variable, some elements in error covariance satisfy an inequality constraint.

In Section 2, the concept of controllability Gramian is introduced. The relationship between error covariance and controllability Gramian is proved. Then a matrix upper bound of error covariance is added. In Section 3.1, some quantitative characteristics of controllability



Gramian, such as the location of peak covariance and the decay rate from peak to bottom, are proved. The tedious algebraic derivations in the proof are included in the Appendix at the end of the paper. In Section 3.2, an algebraic inequality of error covariance deduced from the observation model is proved. In Section 4, we introduce two computational algorithms that are component-based. Section 5 contains three examples demonstrating how to collectively apply the theorems and computational algorithms from previous sections to outline the features of the error covariance.

2. Error covariance and controllability Gramian

Shown in Figure 1 are two error covariance matrices of Kalman filters, one for a linear system of ordinary differential equations and the other for a Discretised shallow water equation. The x - and y - axes represent row and column indices, i and j ; the z -axis represents the absolute value of the covariance, $|P_{ij}|$. In this way, a matrix is visualised as a 3D graph. In this paper, the geometric outline of the 3D graph is referred as the *shape* of error covariance. Both graphs in Figure 1 have similar geometric patterns in which a large number of elements has relatively small value and the covariance has its peaks along one or multiple diagonal lines. Relative to their peak value, both matrices are approximately sparse. In this paper, the shape of error covariance is studied using a set of *quantitative characteristics* that include, but not limited to, the following list of quantifiable geometric properties:

- the peak value and its location,
- the decay rate from peak to bottom,
- inequality constraints of covariance elements.

For high dimensional problems, the true error covariance of a Kalman filter is computationally intractable. In mathematical analysis, a widely used idea in the

estimation of an unknown variable is to find a computationally feasible upper/lower bound. In this section, we prove a relationship between the error covariance of Kalman-Bucy filters and two controllability Gramians. As a corollary of the relationship, a matrix upper bound of the error covariance is found. An advantage of this matrix upper bound is that the quantitative characteristics of its shape are computationally tractable. Let's consider a dynamical system model defined by ordinary differential equations (ODEs)

$$\dot{x}(t) = Ax(t) + qw(t), \quad (1a)$$

$$y(t) = Hx(t) + rv(t) \quad (1b)$$

where $x \in \mathbb{R}^n$ is the state variable, $\dot{x}$ represents its time derivative, $y \in \mathbb{R}^m$ is the observation variable, $w \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ are zero-mean Gaussian white noises with an identity covariance. The matrices $q \in \mathbb{R}^{n \times n}$ and $r \in \mathbb{R}^{m \times m}$ determine the covariances of the model uncertainty and the observation error, respectively. Denote $Q = qq^T$ and $R = rr^T$. Its Kalman-Bucy filter has the following form

$$\dot{\hat{x}}(t) = A\hat{x}(t) + K(t)(y(t) - H\hat{x}(t)), \quad (2a)$$

$$K(t) = P(t)H^T R^{-1} \quad (2b)$$

$$\dot{P}(t) = AP(t) + P(t)A^T + Q - K(t)RK(t)^T. \quad (2c)$$

where $K(t)$ is the Kalman gain, $P(t)$ is the estimation error covariance and $y(t)$ represents observed data. Multiplying e^{-At} and $e^{-A^T t}$ to (2c), we have

$$\begin{aligned} e^{-At} \dot{P}(t) e^{-A^T t} - e^{-At} AP(t) e^{-A^T t} - e^{-At} P(t) A^T e^{-A^T t} \\ = e^{-At} (Q - K(t)RK(t)^T) e^{-A^T t} \end{aligned}$$

This is equivalent to

$$\frac{d}{dt} (e^{-At} P(t) e^{-A^T t}) = e^{-At} (Q - K(t)RK(t)^T) e^{-A^T t}$$

The integration of this equation yields an integral equation of $P(t)$

$$\begin{aligned} P(t) = e^{At} P(0) e^{A^T t} \\ + \int_0^t e^{A(t-\tau)} (Q - K(\tau)RK(\tau)^T) e^{A^T(t-\tau)} d\tau \end{aligned} \quad (3)$$

The integration on the right-hand side requires the history of $K(t)$. If it is integrated in a subinterval $[t_1, t_2] \subseteq [0, t]$, we have a matrix upper bound of the error covariance

$$\begin{aligned} 0 \leq P(t) \leq e^{At} P(0) e^{A^T t} + \int_0^t e^{A(t-\tau)} Q e^{A^T(t-\tau)} d\tau \\ - \int_{t_1}^{t_2} e^{A(t-\tau)} K(\tau) RK(\tau)^T e^{A^T(t-\tau)} d\tau \end{aligned} \quad (4)$$

If $t_1 = t_2$, the second integration equals zero. The matrix upper bound, in this case, represents the error covariance without corrections using observation information. The integrals on the right-hand side of (3) and (4) are closely

related to the concept of controllability Gramian in control theory.

Definition 1. The controllability Gramian of the system

$$\dot{x}(t) = Ax(t) + q(t)u, \quad x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}, u \in \mathbb{R}^p, q \in \mathbb{R}^{n \times p} \quad (5)$$

in the time interval $[t_1, t_2]$ is

$$G^C(t_1, t_2) = \int_{t_1}^{t_2} e^{A(t_2-\tau)} q(\tau) q^T(\tau) e^{A^T(t_2-\tau)} d\tau \quad (6)$$

For the special case $[t_1, t_2] = [0, t]$, we use a simpler notation, $G^C(t)$.

The Gramian, a symmetric positive semidefinite matrix, is a quantitative measure of the sensitivity of system trajectories to the variation of the system's input. If the Gramian has a set of large eigenvalues, then $x(t)$ can reach any given state using a control input that has a relatively small L_2 -norm. The expressions in (3) and (4) are summarised in the following propositions.

Proposition 1. The error covariance, $P(t)$, of the Kalman-Bucy filter (2) satisfies,

$$P(t) = e^{At} P(0) e^{A^T t} + G^C(t) - \hat{G}^C(t) \quad (7)$$

where $G^C(t)$ is the controllability Gramian of (1a) in the time interval $[0, t]$, $\hat{G}^C(t)$ is the controllability Gramian of the control system

$$\dot{\hat{x}}(t) = A\hat{x}(t) + K(t)rv \quad (8)$$

which is the Kalman-Bucy filter (2a) in which $y - H\hat{x}$ is treated as a control input rv .

Proposition 2. The error covariance, $P(t)$, of the Kalman-Bucy filter (2) has an upper bound matrix

$$0 \leq P(t) \leq e^{At} P(0) e^{A^T t} + G^C(t) - \hat{G}^C(t_1, t_2) \quad (9)$$

where $[t_1, t_2] \subseteq [0, t]$. As a special case,

$$0 \leq P(t) \leq e^{At} P(0) e^{A^T t} + G^C(t) \quad (10)$$

Propositions 1 and **2** can be interpreted as follows. In addition to the propagation of $P(0)$, $P(t)$ is determined by the controllability Gramian of the system model subtracting the controllability Gramian of the Kalman filter. If we consider the model uncertainty, w , as a control variable, then its controllability Gramian represents the worst estimation error without any error correction. On the other hand, the controllability Gramian of the Kalman-Bucy filter represents the reduction of error covariance in the filtering process. A more controllable filter results in more aggressive error correction, thus smaller estimation error. Among the three terms in (9), we focus on the the second term, the controllability Gramian of the original system (1). However, the

same computational algorithm that we use for the estimation of G^C is also applicable to the propagation of initial error covariance, the first term in (9). It is summarised in Proposition 6. In this paper, we do not address the computation of the 3rd term in (9). Although this is a very important term that reduces the matrix upper bound using observation model, an efficient computational algorithm for this term in high dimensional spaces has yet to be found.

The inequality (10) has a discrete-time version. Consider the system model,

$$x_k = A_{k-1}x_{k-1} + q_{k-1}w_{k-1}, \quad x_k \in \mathbb{R}^n, A_k \in \mathbb{R}^{n \times n} \quad (11a)$$

$$y_k = H_k x_k + r_k v_k, \quad y_k \in \mathbb{R}^m, H_k \in \mathbb{R}^{m \times n} \quad (11b)$$

where w_k and v_k are serially uncorrelated random variables with a Gaussian distribution that has zero mean and identity covariance. Let $Q_k = q_k q_k^T$ and $R_k = r_k r_k^T$. Its Kalman filter is a discrete-time system in the following form,

$$x_k^- = A_{k-1}x_{k-1}^+, \quad (12a)$$

$$P_k^- = A_{k-1}P_{k-1}^+ A_{k-1}^T + Q_{k-1} \quad (12b)$$

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} \quad (12c)$$

$$x_k^+ = x_k^- + K_k (y_k - H_k x_k^-) \quad (12d)$$

$$P_k^+ = P_k^- - K_k (H_k P_k^- H_k^T + R_k) K_k^T \quad (12e)$$

where x_k^+ and P_k^+ are the analysis state and analysis error covariance, x_k^- and P_k^- are the background state and background error covariance. Using mathematical induction, it is straightforward to prove the following upper bound of P_k^+ .

Proposition 3.

$$P_k^+ \leq A_{k-1} \cdots A_0 P_0^+ A_0^T \cdots A_{k-1}^T + G_k^C \quad (13)$$

where G_k^C is the (discrete-time) controllability Gramian

$$G_k^C = \sum_{j=0}^{k-2} A_{k-1} A_{k-2} \cdots A_{j+1} Q_j A_{j+1}^T \cdots A_{k-2}^T A_{k-1}^T + Q_{k-1} \quad (14)$$

Remark. If the space dimension is high, the controllability Gramian is computationally intractable. However, there exist numerical methods that can compute individual elements in the Gramian, i.e. component-based algorithms. They are introduced in Section 4, including Proposition 5 for linear systems, Proposition 6 for the propagation of the initial error covariance, and (34) for nonlinear systems.

3. Some quantitative characteristics of the shape of error covariance

In this section, we prove a decay rate for $G^C(t)$ and a set of algebraic constraints for the error covariance of Kalman filters.

3.1. The decay rate of $G^C(t)$

Some quantitative characteristics of the shape of $G^C(t)$ can be mathematically proved and numerically computed. The main result in this section is inspired by the approximate sparsity pattern of exponential matrices proved in Iserles (2000). The following function, $c(\alpha, x)$, is used repeatedly in the analysis,

$$c(\alpha, x) = \frac{\alpha^{|x|}}{|x|^{|x|}}, \alpha > 0, x \neq 0. \quad (15)$$

Because $\lim_{|x| \rightarrow \infty} |x|^{1/|x|} = 1$, $\lim_{|x| \rightarrow \infty} \alpha^{1/|x|} = 1$ and $\frac{1}{e} < 1$, we know

$$c(\alpha, x) = \left(\frac{\alpha^{1/|x|}}{|x|^{1/|x|}} \right)^{x^2} > \left(\frac{1}{e} \right)^{x^2}$$

$$c(\alpha, x) = \left(\frac{\alpha}{|x|} \right)^{|x|} < \left(\frac{1}{e} \right)^{|x|}$$

provided that $|x|$ is large enough. This inequality and some properties of $c(\alpha, x)$ are summarised as follows.

Lemma 1. Suppose that α and x are positive numbers. Then $c(\alpha, x)$ is an increasing function of $\alpha > 0$ and a decreasing function of $x \in [\alpha, \infty]$. Furthermore

$$\lim_{x \rightarrow \infty} c(\alpha, x) = 0$$

for any constant α . When $|x| \rightarrow \infty$,

$$e^{-x^2} < c(\alpha, x) < e^{-|x|}. \quad (16)$$

i.e. $c(\alpha, x)$ decreases faster than exponential but slower than the rate of e^{-x^2} .

In the following, a matrix A is said to be banded with bandwidth $s \geq 0$ if $A_{ij} = 0$ whenever $|i - j| > s$. If $s = 0$, then the matrix is diagonal.

Theorem 1. Suppose that A and $Q = qq^T$ in (1) are banded. For any $t > 0$, there exist positive constants $\bar{G}$, α , β , γ , and L so that

$$|(G^C(t))_{ij}| \leq \bar{G} c(\alpha, (|j - i| + \beta)/\gamma), \text{ if } |j - i| > L. \quad (17)$$

Note that the inequality (17) is an upper bound that approximates the decay of G_{ij}^C as $|i - j|$ increases. It implies that $G^C(t)$ peaks around the diagonal in the matrix; and the value of $|G_{ij}^C|$ decreases faster than exponential if moving away from the diagonal. In this theorem, it is assumed that A is banded along the main diagonal. If A consists of multiple banded submatrices, simulations show that the error covariance peaks along corresponding sub-diagonals. In the following, the right-hand side of the equation (17) is called the *decay function*.

The proof of this theorem involves lengthy and tedious algebraic derivations, which is given in Appendix at the end of the paper. In (17), $|i - j| \rightarrow \infty$ implies that the

result is meaningful only if $n \gg L$, where n is the dimension of the state space. In fact, the results in Iserles (2000), the theoretical foundation of Theorem 1, are based upon the assumption that the bandwidth of A satisfies $s \ll n$. Then L is much smaller than n . In Appendix, an analytic estimation of the parameters in (41), i.e. $\bar{G}$, α , β , and γ is proved. However, the estimated value is conservative, which leads to a decay rate much slower than that of the true matrix. In Section 4, a numerical optimisation problem is formulated in (38) that allows to fit the optimal value of the parameters to the component-based estimate of the upper bound matrix (34). Showing in an example, the estimated decay function closely follows the true decay function of $G^C(t)$ as $|i - j|$ increases.

3.2. Some covariance constraints deduced from observation model

An error covariance, as a matrix, satisfies some constraints such as symmetry and positive definiteness. In the following, we prove some covariance constraints resulting from the observation model. In this section, the study is focussed on the discrete-time system (11) and its Kalman Filter (12).

Proposition 4. Suppose the diagonal elements of P_k^+ are bounded by a positive number $P_{\max}$, i.e.

$$|(P_k^+)_{ii}| \leq P_{\max}, \quad 1 \leq i \leq n \quad (18)$$

Then P_k^+ satisfies

$$0 < (H_k P_k^+ H_k^T)_{ii} \leq (R_k)_{ii}, \quad 1 \leq i \leq m, \quad (19a)$$

$$|(H_k P_k^+)_{ij}| < \sqrt{(R_k)_{ii} P_{\max}}, \quad 1 \leq i \leq m, 1 \leq j \leq n, j \neq i \quad (19b)$$

As a special case if the observation measures $(x_k)_i$ for some $1 \leq i \leq n$, then

$$0 < (P_k^+)_{ii} \leq (R_k)_{ii}, \quad (20a)$$

$$|(P_k^+)_{ij}| < \sqrt{(R_k)_{ii} P_{\max}}, \quad 1 \leq j \leq n, j \neq i, \quad (20b)$$

Different from the matrix upper bound in Theorem 1, the algebraic inequalities in Proposition 4 provide upper bounds for the individual elements in an error covariance. For example, if the sensor that measures the i th state variable, $(x_k)_i$, is very accurate, then $(R_k)_{ii}$ is very small. In this case, (20) is a tight upper bound for $(P_k^+)_{ij}$, $1 \leq j \leq n$.

Proof. We prove (20) first. Without loss of generality, let $i = 1$. In this case, the first state variable, $(x_k)_1$, is measured, i.e.

$$H_k = \begin{bmatrix} 1 & 0_{1 \times (n-1)} \\ (H_k)_{2,n,1} & (H_k)_{2,n,2,n} \end{bmatrix}$$

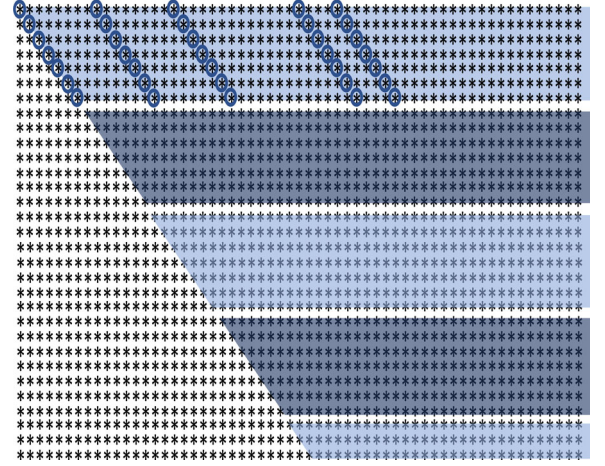


Fig. 2. A partition of error covariance.

where $(H_k)_{2,n,1}$ is the first column of H_k starting from the second row, i.e. $i \geq 2$. The submatrix $(H_k)_{2,n,2,n}$ is similarly defined. Let K_k be the Kalman gain (12c). Define a new gain $\tilde{K}_k$. All rows of $\tilde{K}_k$ equal that of K_k except for the first row, which is $[1 \ 0 \ 0 \ \cdots \ 0]$, i.e.

$$(\tilde{K}_k)_{ij} = (K_k)_{ij}, \quad 1 < i \leq n, 1 \leq j \leq m \quad (21a)$$

$$(\tilde{K}_k)_{1j} = \delta_{1j}, \quad 1 \leq j \leq m \quad (21b)$$

The estimation using $\tilde{K}_k$ is

$$\tilde{x}_k^+ = x_k^- + \tilde{K}_k(y_k - H_k x_k^-)$$

Then $\tilde{x}_k^+$ equals the estimation using the Kalman filter except for the first state,

$$(\tilde{x}_k^+)_{11} = (y_k)_{11}$$

Its error covariance is $(R_k)_{11}$. Because the Kalman gain minimises the trace of P_k^+ (see, for instance, Gelb and The Technical Staff (1974)), we have

$$\begin{aligned} (P_k^+)_{11} &= E((x_k^+)_{11} - (x_k)_{11})^2 \leq E((\tilde{x}_k^+)_{11} - (x_k)_{11})^2 \\ &= (R_k)_{11} \end{aligned}$$

i.e. (20a) holds true. Because P_k is positive definite, the following 2×2 matrix must be positive definite

$$\begin{bmatrix} (P_k^+)_{11} & (P_k^+)_{1j} \\ (P_k^+)_{j1} & (P_k^+)_{jj} \end{bmatrix}.$$

for any $1 \leq j \leq n$. This implies that its determinant is greater than zero,

$$(P_k^+)_{11}(P_k^+)_{jj} - ((P_k^+)_{1j})^2 > 0$$

Therefore,

$$|(P_k^+)_{1j}| < \sqrt{(P_k^+)_{11}(P_k^+)_{jj}}$$

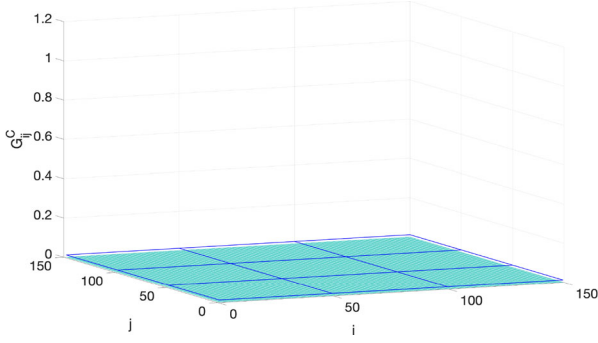


Fig. 3. Rows and columns of error covariance constrained by the observation model.

This inequality implies (20b) because $0 < (P_k^+)_{11} \leq (R_k)_{11}$ and $0 < (P_k^+)_{jj} \leq P_{max}$.

To prove (19), let's focus on $(H_k)_{1,1:n}$, the first row of H_k . The proof for other rows is similar. Because $(H_k)_{1,1:n}$ is not a zero vector, we assume $(H_k)_{11} \neq 0$ (if it is zero, we can re-arrange the index to move a nonzero element to the first position). Define an invertible matrix

$$T = \begin{bmatrix} (H_k)_{11} & (H_k)_{1,2:n} \\ 0_{(n-1) \times 1} & I_{(n-1) \times (n-1)} \end{bmatrix}, \quad (22)$$

$$T^{-1} = \begin{bmatrix} (H_k)_{11}^{-1} & -(H_k)_{11}^{-1}(H_k)_{1,2:n} \\ 0_{(n-1) \times 1} & I_{(n-1) \times (n-1)} \end{bmatrix}.$$

Obviously

$$(H_k)_{1,1:n} T^{-1} = [1 \quad 0 \quad \dots \quad 0]. \quad (23)$$

Using T as a transformation we define a new state variable

$$\tilde{x} = Tx$$

Its dynamical model is

$$\tilde{x}_k = \tilde{A}_{k-1} \tilde{x}_{k-1} + \tilde{q}_{k-1} w_k, \quad \tilde{A}_{k-1} = T A_{k-1} T^{-1}, \quad \tilde{q}_{k-1} = T q_{k-1} \quad (24a)$$

$$y_k = \tilde{H}_k \tilde{x}_k + r_k v_k, \quad \tilde{H}_k = H_k T^{-1} \quad (24b)$$

In $\tilde{H}_k$, the first row is the unit vector (23). Therefore, the associated error covariance satisfies (20). Under the conditional probability distribution $P(\tilde{x}_k | y_{1:k})$, the covariance is

$$\tilde{P}_k^+ = E((\tilde{x}_k - \text{mean}(\tilde{x}_k))(\tilde{x}_k - \text{mean}(\tilde{x}_k))^T) \quad (25a)$$

$$= E(T(x_k - \text{mean}(x_k))(x_k - \text{mean}(x_k))^T T^T) \quad (25b)$$

$$= T P_k^+ T^T \quad (25c)$$

From (20), (22) and (25), we have

$$0 < (H_k P_k^+ H_k^T)_{11} = (\tilde{P}_k^+)_{11} \leq (R_k)_{11},$$

i.e. (19a) holds true. If $j \neq 1$, then

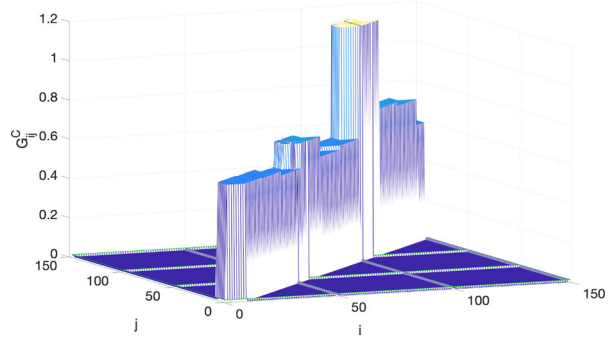


Fig. 4. The approximate peak value of error covariance computed using an upper bound matrix.

$$|(H_k P_k)_{1j}| = (\tilde{P}_k^+)_{1j} < \sqrt{(R_k)_{11} P_{max}}$$

Here, we use the same upper bound P_{max} for both P_k^+ and $\tilde{P}_k^+$ because (22) and (25) imply

$$(P_k^+)_{2:n,2:n} = (\tilde{P}_k^+)_{2:n,2:n}.$$

□

4. Computational issues

The results in previous sections describe various aspects of error covariance. When applied collectively these results together give a global picture of the shape of error covariance. To summarise, we have proved the following quantitative characteristics of the shape of $P(t)$:

- Upper bound

$$P(t) \leq e^{At} P(0) e^{A^T t} + G^C(t), \quad \text{continuous-time} \quad (26a)$$

$$P_k^+ \leq A_{k-1} \dots A_0 P_0^+ A_0^T \dots A_{k-1}^T + G_k^C, \quad \text{discrete-time} \quad (26b)$$

- Decay rate from peak

$$|(G^C(t))_{ij}| \leq \bar{G} c(\alpha, (|j-i| + \beta)/\gamma), \quad \text{if } |i-j| > L(\bar{G}, \alpha, \beta, \gamma, L \text{ are parameters}) \quad (27)$$

- Covariance constraints deduced from observation

If $y_k = (x_k)_i$

$$(P_k^+)_{ii} \leq (R_k)_{ii}, \quad (28a)$$

$$|(P_k^+)_{ij}| < \sqrt{(R_k)_{ii} P_{max}}, \quad 1 \leq j \leq n, j \neq i. \quad (28b)$$

Or in general

$$0 < (H_k P_k^+ H_k^T)_{ii} \leq (R_k)_{ii}, \quad 1 \leq i \leq m, \quad (28c)$$

$$|(H_k P_k^+)_{ij}| < \sqrt{(R_k)_{ii} P_{max}}, \quad 1 \leq i \leq m, 1 \leq j \leq n, j \neq i, \quad (28d)$$

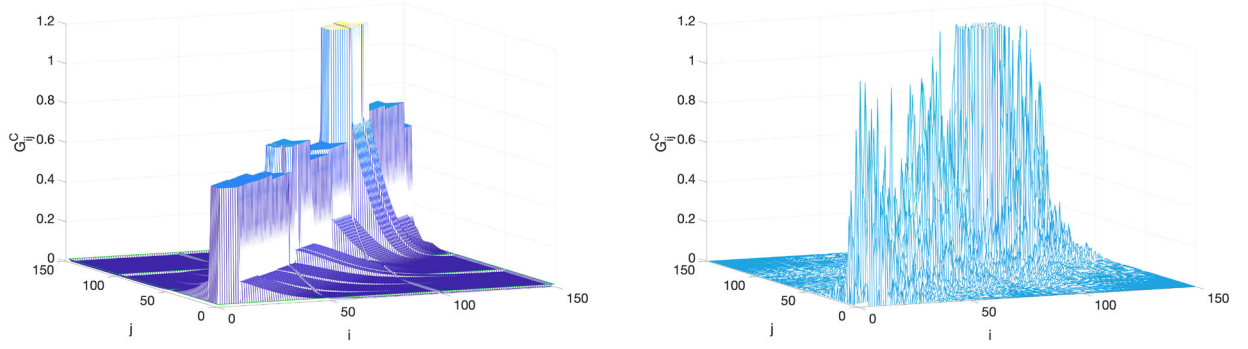


Fig. 5. The approximated shape (left) and the matrix upper bound (right) of the error covariance.

4.1. Computing the elements in $G^C(t)$

Definition 2. The observability Gramian of the system

$$\dot{x}(t) = Ax(t), \quad x \in \mathbb{R}^n, \quad (29a)$$

$$y = H(t)x, \quad y \in \mathbb{R}^m, H \in \mathbb{R}^{m \times n} \quad (29b)$$

at $t = t_1$ in the time interval $[0, t]$ is

$$G^O(t) = \int_0^t e^{A^T(\tau-t_1)} H^T(\tau) H(\tau) e^{A(\tau-t_1)} d\tau \quad (30)$$

The controllability Gramian has the same dimension as the error covariance. Therefore propagating the entire matrix along a trajectory is intractable for problems that have high dimensions. However, the Gramian matrix can be computed element-by-element. In control theory, given a system with state variable x and control input u

$$\dot{x}(t) = Ax(t) + qu, \quad x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}, q \in \mathbb{R}^{n \times n_u}, u \in \mathbb{R}^{n_u}$$

its dual system is a linear system

$$\dot{z}(\tau) = -A^T z(\tau), \quad z \in \mathbb{R}^n \quad (31a)$$

$$y^z(\tau) = q^T z(\tau), \quad y^z \in \mathbb{R}^{n_u} \quad (31b)$$

where z is the dual state, and y^z is the output of the dual system (Kailath, 1980).

Proposition 5. Consider (1a) and its dual system (31).

(i) The controllability Gramian of (1a) in $[0, t]$ equals the observability Gramian of (31) at t , which is

$$G^C(t) = \int_0^t e^{-A(\tau-t)} Q e^{-A^T(\tau-t)} d\tau \quad (32)$$

(ii) Let e_i represent the i th basis vector of $\mathbb{R}^n$. Let $z_i(\tau)$ and $z_j(\tau)$ be the solutions of (31) with final states $z_i(t) = e_i$ and $z_j(t) = e_j$. Then

$$(G^C(t))_{ij} = \int_0^t z_i^T(\tau) Q z_j(\tau) d\tau, \quad (33)$$

The formula (33) implies that individual elements in G^C can be computed without the need to evaluate the entire matrix, provided that the dual system, or the

adjoint model, can be numerically propagated reversely in time. Integrating the adjoint model twice, $z_i(\tau)$ and $z_j(\tau)$, one can compute three elements in the matrix, $(G^C(t))_{ii}$, $(G^C(t))_{jj}$ and $(G^C(t))_{ij} = (G^C(t))_{ji}$.

Proof. Based on Definitions 1 - 2, Part (i) is obvious. To prove (ii), Note that $z_i(\tau) = (e^{-A^T(\tau-t)})e_i$, where e_i denotes the i th basis vector. Therefore, the i th row and j th column of $G^C(t)$ is the integration of the inner product of $q^T z_i(\tau)$ and $q^T z_j(\tau)$. Therefore, (33) holds, where $Q(\tau) = q(\tau)q^T(\tau)$. $\square$

The idea used in Proposition 5 is also applicable to the propagation of initial error covariance.

Proposition 6. Suppose B_0 is a matrix satisfying

$$P(0) = B_0 B_0^T$$

Consider the following dual system

$$\dot{z}(\tau) = A^T z(\tau), \quad z \in \mathbb{R}^n$$

$$y^z(\tau) = B_0^T z(\tau), \quad y^z \in \mathbb{R}^n$$

Let $z_i(\tau)$ and $z_j(\tau)$ be the solutions satisfying $z_i(0) = e_i$ and $z_j(0) = e_j$. Then

$$\begin{aligned} (e^{A^T t} P(0) e^{A^T t})_{ij} &= \int_0^t z_i^T(\tau) B_0 B_0^T z_j(\tau) d\tau, \\ &= \int_0^t z_i^T(\tau) P(0) z_j(\tau) d\tau \end{aligned}$$

Propositions 5 and 6 are proved for continuous-time linear systems. To derive its discrete-time version, let Δt be the step size, $x_k = x(k\Delta t)$, $t = K\Delta t$ for some positive integer K . A Discretisation of (1a) is

$$x_k = e^{A\Delta t} x_{k-1} + \bar{q} v_k, \quad \bar{q} = \sqrt{\Delta t} q$$

Define

$$z_s(k) = e^{(K-k)A\Delta t} e_s, \quad s = i, j$$

Denote $\bar{q}^T \bar{q}$ by $\bar{Q}$. The Euler Discretisation of (33) is

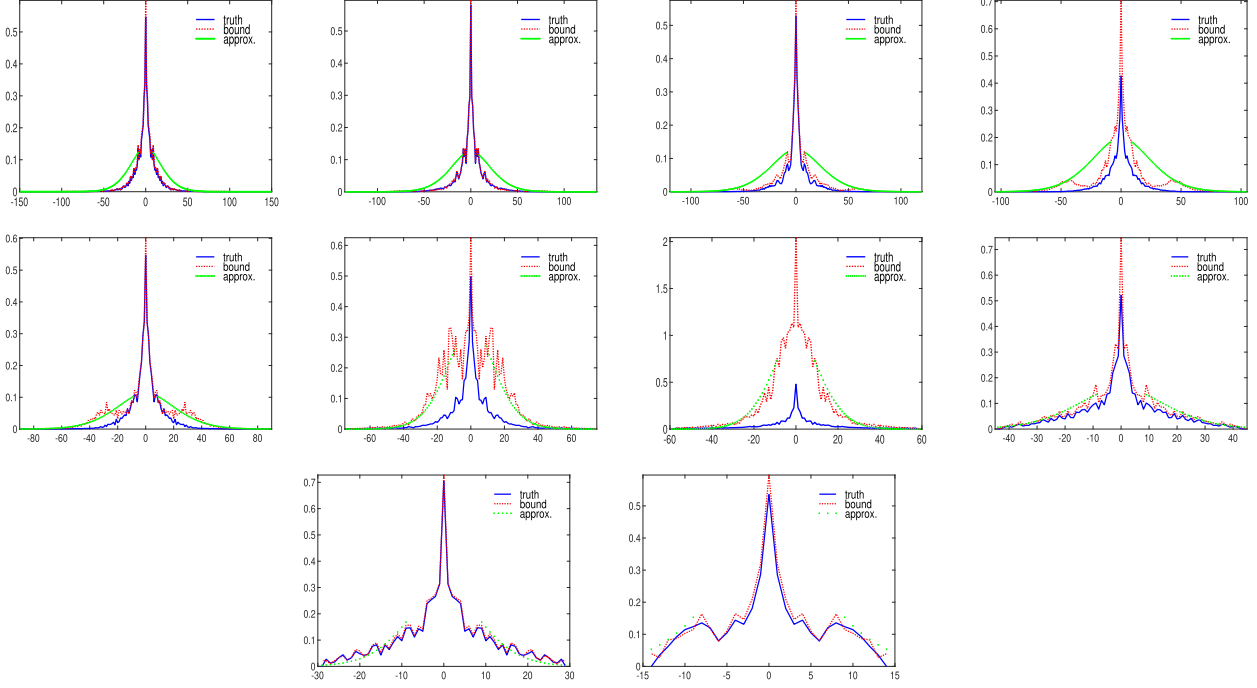


Fig. 6. The averaged error covariance along diagonals of blocks. The horizontal axis represents the indices of diagonals, where the middle point is the main diagonal. The vertical axis represents the average value of the elements in the error covariance along diagonals. The blue curve represents the true error covariance of a linear Kalman filter. The red curve is the matrix upper bound of error covariance. The green curve represents the approximated decay function.

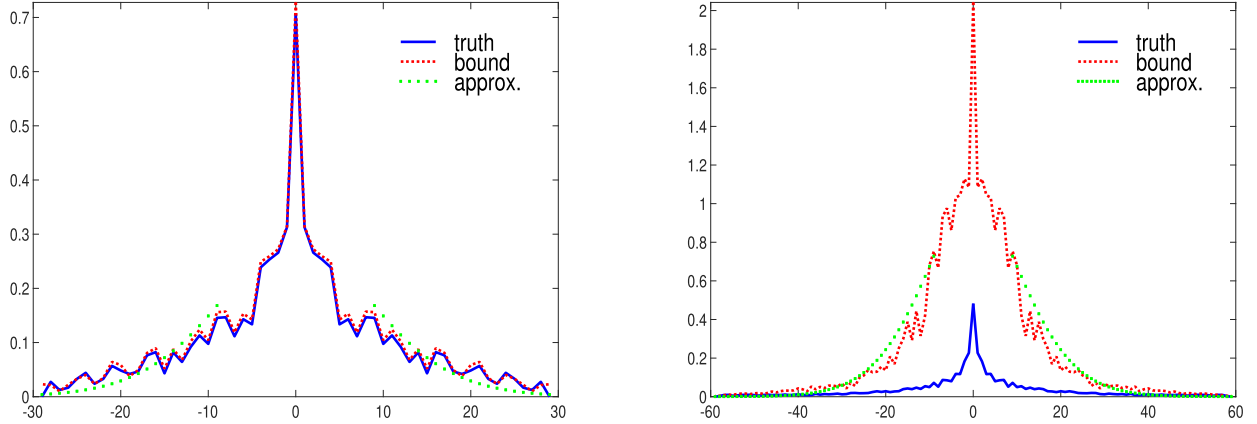


Fig. 7. The averaged error covariance along diagonals of two representative blocks. See the caption of Figure 6 for details.

$$G_{ij}^C = \sum_{k=1}^{K-1} z_i^T(k) \bar{Q} z_j(k) + \bar{Q}_{ij} \quad (34)$$

For a nonlinear system

$$x_k = M(x_{k-1}) + \bar{q}v_k$$

Given a trajectory $\{x_k\}_{k=0}^K$. Let M_k^* represent the adjoint model at x_k . Then (34) is an approximation of controllability Gramian in which

$$z_s(k) = M_k^* \circ M_{k+1}^* \cdots \circ M_{K-1}^* e_s, \quad s = i, j \quad (35)$$

This nonlinear version is a form of empirical Gramians. It is used in various applications of nonlinear systems, from ocean dynamics, power systems, to WRF data assimilation (Kang and Xu, 2009; Krener and Ide, 2009; Kang and Xu, 2012; Yoshimura et al., 2020).

4.2. Decay rate approximation

In the next, we introduce a method of approximating the parameters, $(\bar{G}, \alpha, \beta, \gamma)$, in the decay function (27). Their

value given in the proof of [Theorem 1](#) (Appendix) is conservative, which results in a decay rate that is much slower than that of G^C . If one can compute the value of a subset of elements in $G^C(t)$, which is possible as proved in [Proposition 5](#), then we can find $(\bar{G}, \alpha, \beta, \gamma)$ by using curve fitting. More specifically, assume that the set of elements

$$\{G_{ij}^C | (i, j) \in \mathcal{I}\} \quad (36)$$

is known, where $\mathcal{I}$ is an index subset for $i, j \in \{1, 2, \dots, n\}$. It is easy to check that $c(\alpha, x)$ reaches its maximum value at $x = \alpha/e$. Set

$$\beta = \frac{\alpha\gamma}{e} \quad (37)$$

Then we compute the following minimisation

$$(\bar{G}, \alpha, \gamma) = \arg \min_{\bar{G}, \alpha, \gamma} \sum_{(i,j) \in \mathcal{I}} \left(\bar{G}c(\alpha, \alpha/e + (|j-i|)/\gamma) - |G_{ij}^C| \right)^2 \quad (38)$$

In practical applications, the location of $(i, j) \in \mathcal{I}$ is important. We recommend selecting the set $\mathcal{I}$ so that the points are distributed with some close to the peak and some away from the peak. In the examples in [Section 5](#), $\mathcal{I}$ contains five diagonals ([Figure 2](#)) for each block. It covers the range from the peak (along main diagonal in the examples) to the area where the elements of error covariance are very small (the tail of the decay function).

4.3. Component-based computation

Because the elements in G^C can be computed individually using (33) or its Discretisation (34), one can focus on a given area or a block in a matrix to find its shape without the need to know the covariance value outside the block, thus the term ‘component-based.’ Three examples

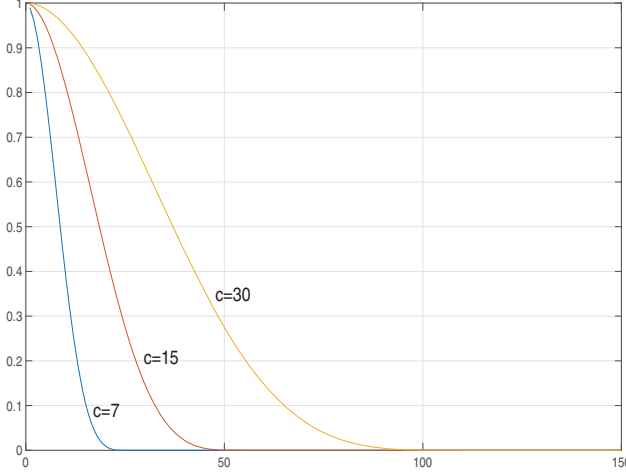


Fig. 8. Correlation functions.

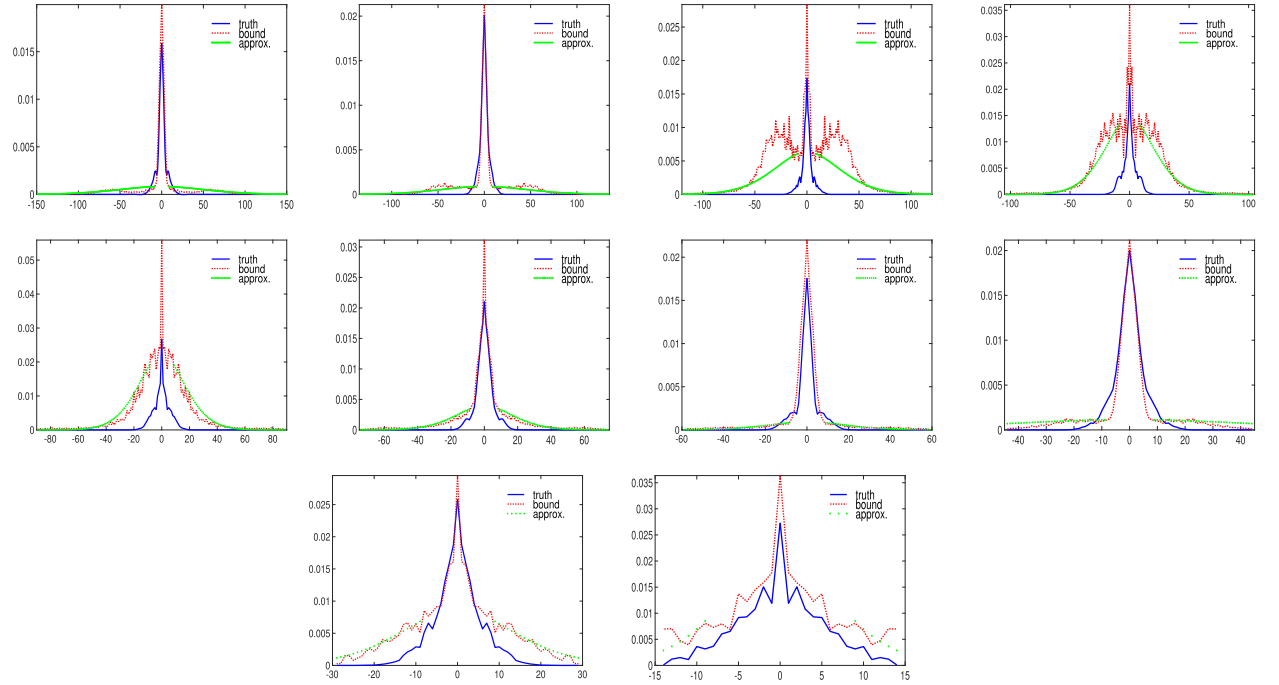


Fig. 9. The averaged error covariance along diagonals of blocks for localisation $c = 30$. The horizontal axis represents the indices of diagonals, where the middle point is the main diagonal. The vertical axis represents the average value of the elements of error covariance along diagonals. The blue curve represents the error covariance of an EnKF. The red curve is the matrix upper bound of error covariance. The green curve represents the approximated decay function.

are given in Section 5. The main computational steps are summarised as follows.

1. *Partition* a large matrix into blocks shown in Figure 2. The elements in a matrix are represented by “*” in the plot. A partition of the matrix consists of blocks of elements, which are represented by shaded areas with different colours. The partition is shown in the figure for the upper half of the matrix. The lower half has the same value due to the symmetry of error covariance.
2. *The peak value* in each block is the average of its diagonal. The computation of diagonal elements is based on (34), the Discretisation of (33).

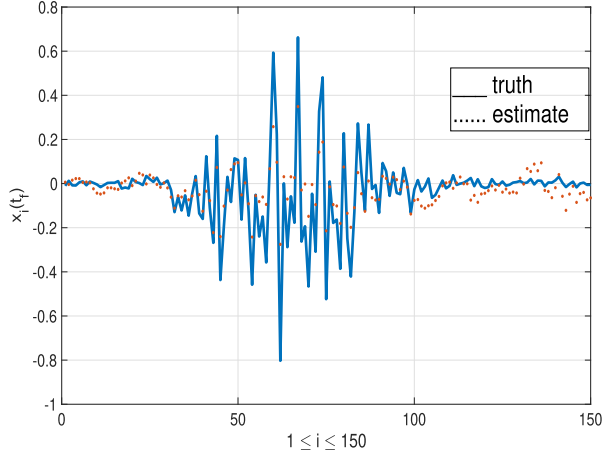


Fig. 10. The state $x_1(t_f), x_2(t_f), \dots, x_n(t_f)$ where $n = 150$. Solid: truth. Dotted: estimated value.

3. *Decay function* is estimated using (38). To apply this formula of curve fitting, a set of elements in G^C (represented by circles in Figure 2) is computed using (34).

Note that the results in Section 3.2, i.e. covariance constraints (28) deduced from observation model, are not included in these steps. The constraints can be applied to individual rows/columns separately in addition to the computed decay rate. It tends to be a more accurate upper bound but applicable to only special rows determined by the observation model.

5. Examples

In the following, we give three examples in which the quantitative characteristics in Section 3 are computed to approximate the upper bound matrix of the error covariance. The outline shapes of the upper bound matrix and the true error covariance are shown in plots. In the first example we demonstrate the computational algorithm. In the second example, the approximated outline shape of an upper bound matrix is used as an indicator of upper bound in the determination of localisation radius. The third example is a case study in which we explore the application of the results proved in previous sections to a nonlinear system that models a tsunami wave using the shallow water equations.

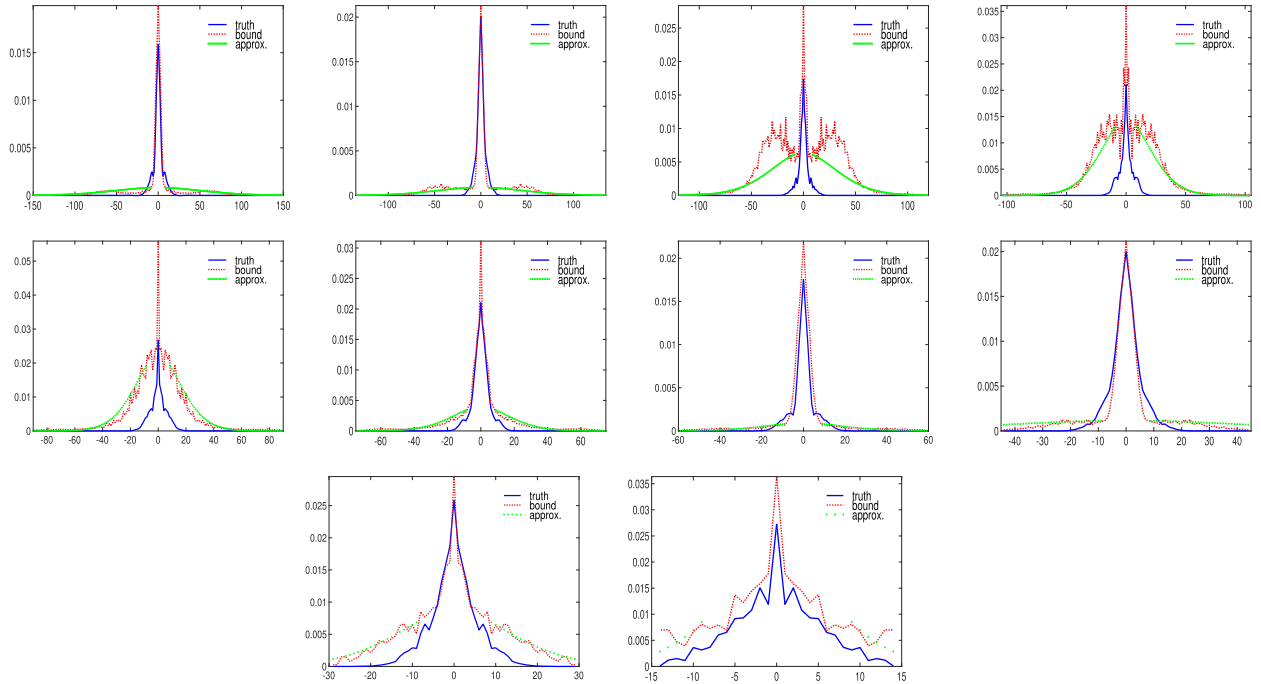


Fig. 11. The averaged error covariance along diagonals of blocks for localisation $c = 7$. See the caption of Figure 9 for details.

5.1. Example 1

The illustrative example has $n=150$ state variables. This dimension is chosen because it is small enough so that the Kalman filter is computationally tractable; and it is

about 2% of the standard deviation of the initial state. The observation measures x_1 , x_{50} , x_{100} , and x_{150} . The initial error covariance is $P(0) = Q$, i.e. the only known information about the background error is the model uncertainty. In the simulations, the final time is $t_f = 10$.

$$\begin{aligned}
 A &= \begin{bmatrix} -2.4689 & -0.7236 & 0.1301 & -0.2498 & -0.2450 & \cdots & \cdots \\ 0.7227 & -2.2126 & 0.7261 & 0.6476 & 0.6578 & \cdots & \cdots \\ 0.4485 & -0.0722 & -2.7618 & -0.1565 & -0.7478 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \begin{pmatrix} \text{random value} \\ \text{banded matrix} \\ \text{bandwidth} = 8 \end{pmatrix} \\
 q = \sigma_Q &= \begin{bmatrix} 1.00 & 1.00 & 0.50 & 0.33 & 0.25 & 0 & 0 & 0 & \cdots & \cdots \\ 1.00 & 1.00 & 1.00 & 0.50 & 0.33 & 0.25 & 0 & 0 & \cdots & \cdots \\ 0.50 & 1.00 & 1.00 & 1.00 & 0.50 & 0.33 & 0.25 & 0 & \cdots & \cdots \\ 0.33 & 0.50 & 1.00 & 1.00 & 1.00 & 0.50 & 0.33 & 0.25 & \cdots & \cdots \\ 0.25 & 0.33 & 0.50 & 1.00 & 1.00 & 1.00 & 0.50 & 0.33 & \cdots & \cdots \\ 0 & 0.25 & 0.33 & 0.50 & 1.00 & 1.00 & 1.00 & 0.50 & \cdots & \cdots \\ 0 & 0 & 0.25 & 0.33 & 0.50 & 1.00 & 1.00 & 1.00 & \cdots & \cdots \\ 0 & 0 & 0 & 0.25 & 0.33 & 0.50 & 1.00 & 1.00 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix}, \quad \sigma_Q = 0.2121 \\
 y &= [x_1 \quad x_{50} \quad x_{100} \quad x_{150}] + rv(t), \\
 r &= \begin{bmatrix} 0.0102 & 0.0098 & 0.0044 & 0.0110 \\ -0.0083 & -0.0088 & 0.0028 & -0.0014 \\ -0.0097 & 0.0034 & -0.0103 & -0.0035 \\ 0.0047 & -0.0080 & 0.0067 & -0.0102 \end{bmatrix}
 \end{aligned} \tag{39}$$

large enough so that one can demonstrate the sparsity of error covariance. It is defined in the form of differential equation (1). The matrices are partially shown in (39). The A matrix is banded whose bandwidth is $s=8$. For many methods of PDE Discretisation, the resulting ODE system has a bandwidth less than $s=8$. To generate the elements in A , we first form a matrix using random numbers with the uniform distribution in $[-1, 1]$. Then we add a diagonal matrix aI to it, where a is chosen to make sure the system is stable, i.e. placing the real parts of all eigenvalues less than zero. Assuming that model uncertainty has its impact in a local area only, the matrix q is also banded in which $s=4$; thus the bandwidth of $Q = qq^T$ is 8. The matrix associated with observation error is r . We assume a sensor sampling rate $dt=0.125$. The elements of r in the Discretised model are random numbers uniformly distributed in $[-0.0113, 0.0113]$. This range is

The σ_Q value above is equivalent to continuous-time white noise of model uncertainty with a standard deviation $\sigma=0.6$. The component-based computational method makes it possible to focus on any given area in the covariance matrix. In this example we divide the matrix into ten blocks. Each block is a strip of rows above the diagonal of the matrix (Figure 2). The goal is to approximate the shape of the error covariance in each block. The number of blocks in the partition depends on various factors. Basically, increasing the number of blocks refines the approximation with the cost of higher computational load. As the first step, we apply the inequalities (28) deduced from the observation model to this problem. More specifically, we have

$$\begin{aligned}
 (P_k^+)_{ii} &\leq (R_k)_{ii}, \quad i = 1, 50, 100, 150 \\
 |(P_k^+)_{ij}| &< \sqrt{(R_k)_{ii} P_{max}}, \quad i = 1, 50, 100, 150, j \neq i, 1 \leq j \leq n
 \end{aligned}$$

A guaranteed upper bound, P_{max} , is not available. Using component-based computation based on the dual system, we compute the diagonal elements of G^C to find that most of them satisfy $G_{ii}^C < 1$. It indicates that $P_{max} \approx 1$. So, the upper bound of error covariance along these four rows and columns can be determined, shown in Figure 3.

In the next step, we approximate the peak value of each block. We use the peak value of G^C as an approximation because the impact of the initial error decays rapidly for the stable system. The computation is based on (34), the Discretisation of (33). Then the average value is used to represent the peak value over each block. This is shown in Figure 4.

Finally, we estimate the decay rate by finding the decay function $c(\alpha, x)$ of the matrix upper bound. For this purpose, it is necessary to evaluate a subset of elements for each component. As shown in Figure 2, five diagonals are evaluated for each block. Then, the solution of (38) based on the computed $|G_{ij}^C|$ determines the parameter value in the decay function $c(\alpha, x)$. The computations are based on the absolute value of entries in G^C . The approximated shape of error covariance is shown in Figure 5 (left). For comparison, the plot of the matrix upper bound is also shown in Figure 5 (right). The approximate shape captures the key feature, such as peak and decay, very well.

For a more detailed comparison, the average value along the main diagonal as well as all other diagonals are shown in Figure 6 for every block. All computations are based on the absolute value of entries in covariance matrices. There are a total of ten blocks. The gap between the true error covariance and its matrix upper bound varies. Shown in Figure 7 are two enlarged plots from Figure 6. The plot on the left shows that the matrix upper bound (red) is quite close to the true error covariance (blue); and the approximated decay curve (green) is

right shows a block in which the Kalman filter is very effective in making error corrections. As a result, the true error covariance is much smaller than the matrix upper bound. In this block, the matrix upper bound over estimates the true error covariance by a large amount. For improvement, one needs to compute the third term in (9), a problem that cannot be solved using the results in this paper. To summarise, the matrices of error covariance and its upper bound are computationally intractable for high dimensional problems. However, using the component-based algorithm, the decay function of the error covariance upper bound can be numerically computed over a given region.

5.2. Example 2

In applications of EnKFs, finding a localisation radius is essential to the filter's accuracy. Because an upper bound matrix decays exponentially when $|i - j|$ increases (Theorem 1), this upper bound provides valuable information about the localisation radius. The matrices, A , q , and r in this example are generated in the same way as in Example 1 with some parameter value changes. More specifically, σ_Q and r in (39) are assigned the following value

$$\sigma_Q = 0.0707$$

$$r = \begin{bmatrix} 0.0917 & -0.1071 & -0.0122 & -0.0239 \\ -0.0992 & 0.0014 & 0.0082 & -0.1032 \\ -0.0341 & -0.0700 & -0.0441 & -0.0586 \\ 0.0330 & -0.0524 & 0.0995 & 0.0297 \end{bmatrix}$$

Similar to Example 1, the observation measures x_1, x_{50}, x_{100} and x_{150} . For the estimation of $x(t)$, we use an EnKF in which the ensemble size is $n_{ens} = 15$. The localisation is based on the correlation function as defined in Gaspari and Cohn (1999)

$$\rho = \begin{cases} -\frac{1}{4}(|z|/c)^5 + \frac{1}{2}(|z|/c)^4 + \frac{5}{8}(|z|/c)^3 - \frac{5}{3}(|z|/c)^2 + 1, & 0 \leq |z| \leq c, \\ \frac{1}{12}(|z|/c)^5 - \frac{1}{2}(|z|/c)^4 + \frac{5}{8}(|z|/c)^3 + \frac{5}{3}(|z|/c)^2, & c \leq |z| \leq 2c, \\ -5(|z|/c) + 4 - \frac{2}{3}(c/|z|), & 2c \leq |z|. \\ 0, & \end{cases}$$

very close to both. In this block, the approximated shape is very close to the true shape of the error covariance. However, this is not always the case. The plot on the

The value of c determines the localisation radius. The curves in Figure 8 are three correlation functions for $c = 7, 15, 30$.

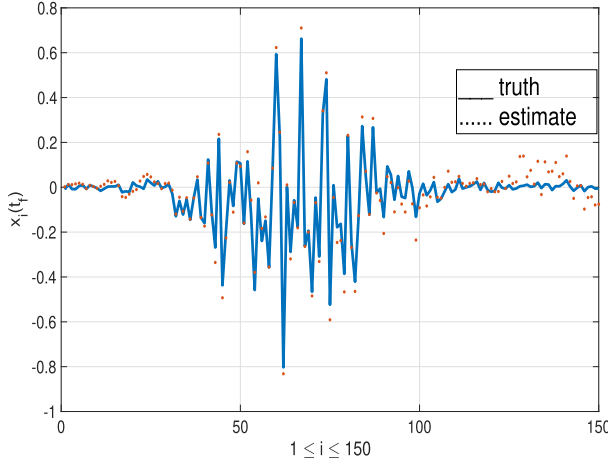


Fig. 12. The state $x_1(t_f), x_2(t_f), \dots, x_n(t_f)$ where $n=150$. Solid: truth. Dotted: estimated value.

In the simulation, we first apply the localisation with the largest radius, $c=30$. Following the same approach in Example 1, we divide the covariance matrix into ten blocks. Each block is a strip of rows above the diagonal of the matrix (Figure 2). The matrix upper bound over each block at $t_f = 10$ is computed and then compared to the average value of the EnKF error covariance along diagonals shown in Figure 2 (taking absolute value before averaging). In Figure 9, the horizontal axis in each plot represents the indices of diagonals, where the middle point is the main diagonal. The vertical axis of each plot represents the average value of the elements of error covariance along diagonals. The blue curve is the error covariance of an EnKF. The red curve is the matrix upper bound of error covariance. For problems with high dimensions, both matrices are computationally intractable. In this case, what can be computed is the approximated decay function. It is shown as the green curve in the plots. In Figure 9, Plot(1,1) (the plot located in the first row first column), Plot(1,2), Plot(2,3) and Plot(2,4) show that the average value of error covariance of the EnKF is much larger than the decay function, implying that the localisation radius is too large. Figure 10 shows the state and its EnKF estimation at $t = t_f$, where $t_f = 3.75$. The estimation missed all peaks located in the middle part of the plot. The relative error is greater than 0.51.

Using the correlation function with $c=15$, the radius of localisation is much smaller. The computation shows that the average value of error covariance is below or close to the decay function in eight of the ten blocks. However, there are two blocks where the average value of error covariance is still much larger when it is compared to the decay function. Comparing to the previous case of $c=30$, the estimated value of $x(t_f)$, the final state, follows the peaks of the truth, although the filter

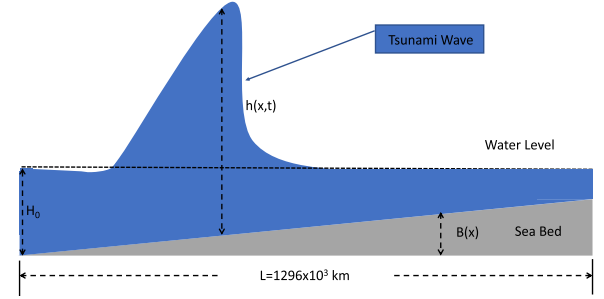


Fig. 13. The variables in the model of a tsunami wave.

overestimates the value at some points. The overall relative error is reduced to 0.47.

If $c=7$, the localisation has the smallest radius among the three correlation functions in Figure 8. In Figure 11, all blue curves are below or close to the green line, implying that the radius of localisation is bounded by the decay function. The final state $x(t_f)$ and its estimation is shown in Figure 12. It is obviously improved relative to the previous cases when the location radius is too large. In fact, the relative estimation error is less than 0.25.

Remark. In this example, the decay function provides an indicator of upper bound in the determination of localisation radius. From (9), the radius of localisation should be the radius of G^C with a correction due to the term $\hat{G}^C(t_1, t_2)$ (for the simplicity in argument we assume that the impact of the initial error covariance is small). This example shows that such correction reduces the radius. Elements of the error covariance matrix resulting from a sound localisation radius are expected to lie under the estimate of the matrix upper bound.

5.3. Example 3 - the shallow water equations

The theorems proved in this paper are based on a common assumption: the system is linear. Generalising the theory to nonlinear systems is beyond the scope of this paper. Instead, the following is a case study in which we explore the application of the results proved in previous sections to a nonlinear system. It is an example adopted from Deleanu and Dumitrache (2019) in which a tsunami wave is simulated using the shallow water equations. Illustrated in Figure 13, the water wave travels a horizontal distance of $L = 1,296 \times 10^3$ m. It approaches the shore on a variable seabed depth that has a constant slope $s = 40/L$. The wave propagation is modelled using the shallow water equations

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(uh) = 0 \quad (40a)$$

$$\frac{\partial}{\partial t}(uh) + \frac{\partial}{\partial x}\left(v^2h + \frac{1}{2}gh^2\right) = -gh\frac{dB}{dx}. \quad (40b)$$

where $h = h(t, x)$ is the flow depth, $u = u(t, x)$ is the flow velocity, $B(x)$ is the seabed height, g is the gravitational acceleration and $(t, x) \in [0, T] \times [0, L]$. In this example, the wave propagation is simulated for the time interval $T = 60,000$ s. The result is used as the truth in the study of error covariance. A total of $m = 80$ sensors are placed along the seabed with an equal distance of 16,200 m apart. Each sensor measures the full state $[h, uh]^T$. The value of parameters is summarised in Table 1 (also see Figure 13 for illustration).

The boundary conditions are

$$\begin{aligned} B(x) &= \frac{40x}{L} \\ h(t, 0) &= 64.5 + 3 \sin \left(\pi \left(\frac{4t}{86400} - \frac{1}{2} \right) \right), \quad h(t, L) = H_0, \\ u(t, 0) &= \sqrt{g} \left(\sqrt{h(t, 0)} - \sqrt{H_0} \right), \quad u(t, L) = 0 \\ h(0, x) &= H_0, \quad u(0, x) = 0 \end{aligned}$$

Coded in Matlab, the shallow water equations are numerically solved using the first order Lax-Friedrichs algorithm (LeVeque, 1992) over equally spaced grid points

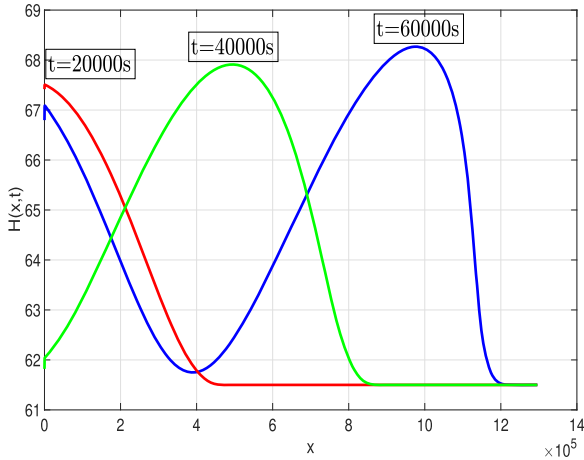
$$\begin{aligned} x_0 &= 0 < x_1 < x_2 < \dots < x_{N_x} = L, \\ t_0 &= 0 < t_1 < t_2 < \dots < t_{N_t} \end{aligned}$$

In the Discretisation, $\Delta x = 324$ m, $N_x = 4,001$, $\Delta t = 10$ s and $N_t = 6,000$. The associated tangent linear model and adjoint model are also programmed in Matlab. Shown in Figure 14 is the nominal trajectory around which the error covariance is analysed.

An error covariance in full dimension, 8002×8002 , is computed using unscented Kalman Filter (UKF) (Julier

Table 1. The parameters in Example 3.

Parameter	L	T	H_0	g
Value	$1,296 \times 10^3$ m	60,000 s	61.5m	9.81 m/s^2



et al., 2000; Julier and Uhlmann, 2004). The model uncertainty and the observation noise are introduced in the same way as that in (1),

$$\begin{aligned} z_k &= M(z_{k-1}) + qw_{k-1} \\ y_k &= H(x_k) + rv_k \end{aligned}$$

where $M(x)$ represents the Discretised shallow water equations and

$$q = \begin{bmatrix} 0.04 & 0.04 & 0.03 & 0.02 & 0.01 & 0 & \dots & 0 \\ 0.04 & 0.04 & 0.04 & 0.03 & 0.02 & 0.01 & \dots & 0 \\ 0.03 & 0.04 & 0.04 & 0.04 & 0.03 & 0.02 & \dots & 0 \\ 0.02 & 0.03 & 0.04 & 0.04 & 0.04 & 0.03 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix},$$

$$r = \begin{bmatrix} 0.03 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0.03 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0.03 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0.03 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

Obviously, q is a banded matrix of bandwidth $s=4$ and r is diagonal. Shown in Figure 15 is the error covariance of the UKF. Different from linear systems, the estimation from a UKF is not guaranteed to be optimal. However, UKF has been a widely used nonlinear filter known for its good accuracy and robust stability. In this example, it serves as a reference for comparison. The water wave and its estimation at $t = 60,000$ s is shown in Figure 16. The relative error of $u(t, x)$ is 0.050 and the relative error of $h(t, x)$ is 0.005. For the component-based computation of G^C , the adjoint model is applied in (34)–(35). The length of time window is K . We choose a value of K different from N_t . Its value should be large enough so that the initial error is diminished. On the other hand, we do not want to apply a K that is too large because the adjoint model is a linearisation that introduces additional error

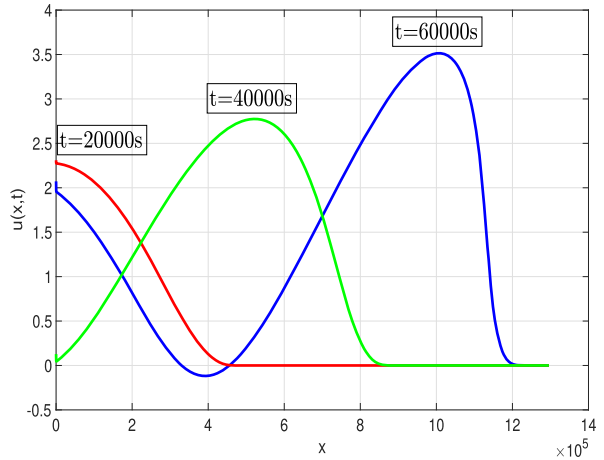


Fig. 14. Water wave at $t = 20,000, 40,000, 60,000$ s.

in each iteration. In this example $K=150$, a time window during which the UKF reduces about 75% of the initial error. Following the same approach in the previous examples, we divide the covariance into blocks as shown in Figure 2. For illustrative purpose, let us focus on the block between the rows $2001 \leq i \leq 2100$. The average value along diagonals in this block is shown in Figure 17. The figure contains three plots because the error covariance as a matrix has three submatrices, odd row vs odd column (covariance of h), even row vs even column (covariance of uh), and odd row vs even column (covariance between h and uh).

The goal of this example is to explore if the results proved for linear systems are applicable to a nonlinear system such as the shallow water equations. We first check the results in Proposition 2 and Theorem 1. In other words, we want to check if the controllability Gramian G^C provides a matrix upper bound; and if it decays at the rate that is claimed in Theorem 1. From Figure 17, the average value of G_{ij}^C (red) is located on top

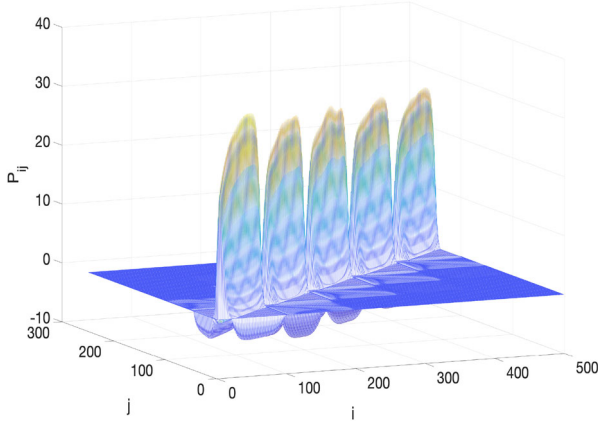


Fig. 15. UKF error covariance (the portion $1 \leq i, j \leq 500$).

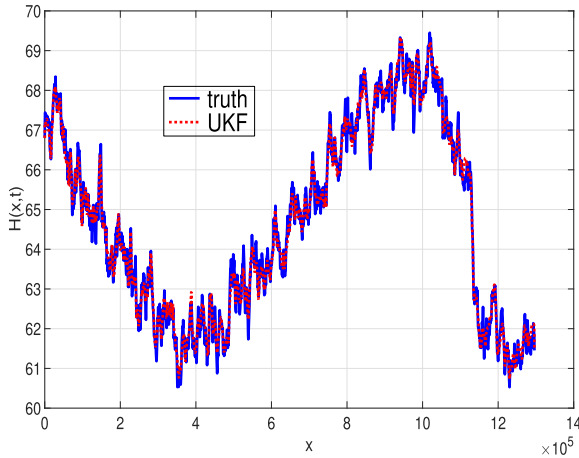
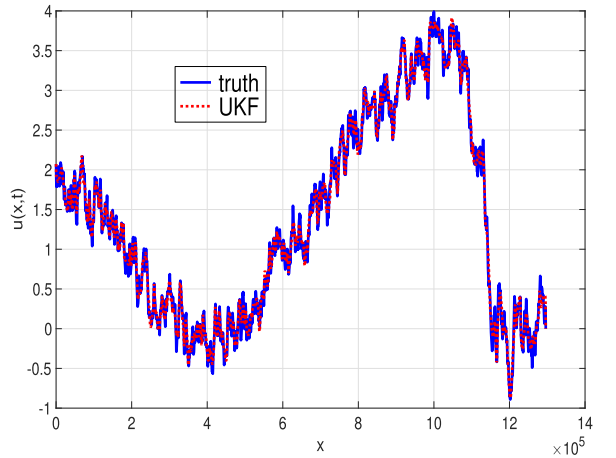


Fig. 16. UKF estimation at $t=60,000$ s. Solid: truth. Dotted: estimated value.

of the average P_{ij} (blue). It implies that G^C is an upper bound of the error covariance. A decay function (green curve) is plotted as an approximation of G^C . It is computed based on four diagonals of G^C in the block. For this nonlinear example, it shows that the decay curve agrees with G^C (red) around the diagonal when the covariance is high. However, different from the linear examples, G^C approaches zero slowly when the location is moving away from the diagonal. Nevertheless, the decay curve approaches zero anyway. As a result, the decay curve serves as a better estimate of the error covariance than G^C . It is unclear if this is a general phenomenon or just a special case, a topic for further research.

In the next, we verify the results in Proposition 4, more specifically the inequalities in (20). It is a set of upper bounds determined by the observation model. As an example, we focus on the row $i=2001$ and $i=2002$ in P_k^+ . The corresponding state variables, x_i , are directly measured by a sensor. In Figure 18, the green line on top is the constant on the right side of the inequality (20b). The blue curve below it is the value of elements in the i th row of P_k^+ (only the columns $1850 \leq j \leq 2150$ are shown here). Clearly, the inequality (20b) holds for this nonlinear example. The red mark at $j=2001$ in the left plot and $j=2002$ in the right plot in Figure 18 is R_{ii} , the right side of the inequality (20a). It is below the blue curve, which implies that R_{ii} is not an upper bound of $(P_k^+)_{ii}$. Therefore, (20a) does not hold for this nonlinear example. The difference is $(P_k^+)_{ii} - R_{ii} = 0.0075$ for $i=2001$ and 0.0076 for $i=2002$. In fact, this difference is relatively insignificant because the covariance at grid points without a sensor has much larger value. For instance, the row adjacent to $i=2002$ in P_k^+ is shown in Figure 19 (green curve). Relative to it, the i th row (blue curve) is so small the different between R_{ii} and $(P_k^+)_{ii}$ is



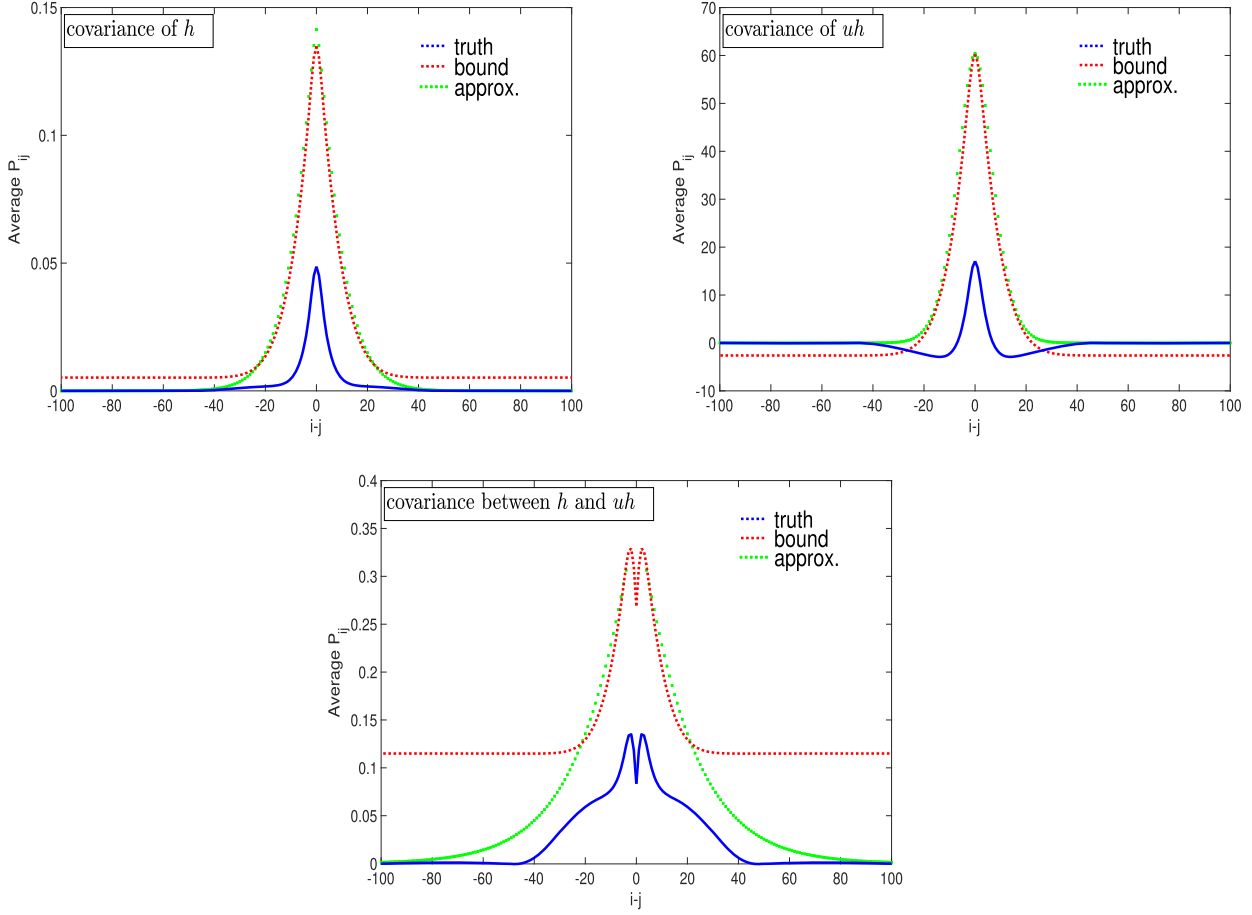


Fig. 17. The averaged error covariance along diagonals of the block between the rows $2001 \leq i \leq 2100$. See the caption of Figure 6 for details.

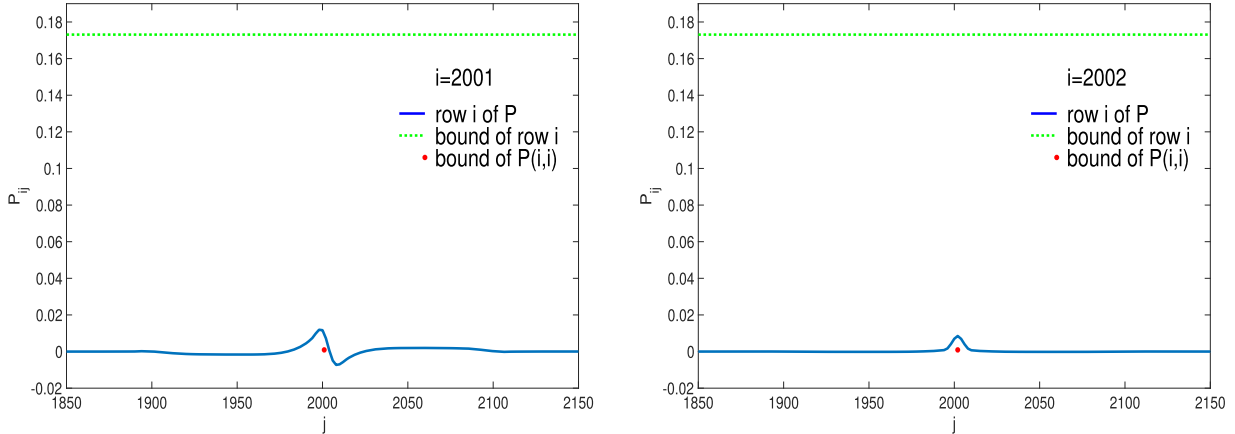


Fig. 18. The upper bound of error covariance deduced from observation model.

negligible. In other words, even though the inequality (20a) does not hold, R_{ii} still provides a good approximation of $(P_k^+)_{ii}$ with a small relative error. The observation model in this example is linear. We speculate

that the reason why the inequality does not hold is because of the nonlinear dynamics of the system. The deep investigation of this behaviour is the subject of an on-going study.

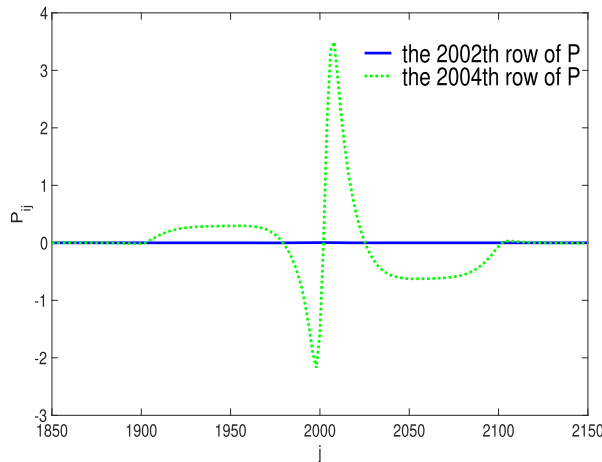


Fig. 19. Error covariance along two adjacent rows, $i=2002$ and $i=2004$.

6. Conclusions

Several quantitative characteristics are introduced and studied that can be used to outline the shape of error covariance as a 3D graph. The proved quantitative characteristics include matrix upper bounds of error covariance, the decay rate from peak to bottom of an upper bound matrix, and some inequality constraints of error covariance deduced from observation models. The concept of controllability Gramian and its computational methods play an important role in several parts of the study. These unveiled interconnections between error covariance and system models are new discoveries in the literature of Kalman filters. Although the computation of error covariance is generally intractable for high dimensional systems, computational methods developed in this paper can numerically compute the quantitative characteristics with a limited computational load. Some ideas of using the results to improve Kalman filter applications are explored in three examples, in which more questions are raised for future research than that are solved. For instance, the quantitative characteristics are collectively, but not interactively, applied. Specifically, the computational algorithm of the decay function does not use the constraints deduced from the observation model. The constraints are applied to the approximated shape of error covariance after the decay function is computed. In addition, the controllability Gramian of the Kalman-Bucy filter in (9) is a correction term. Based on the observation model, this term reduces the error of the matrix upper bound. As a problem for future research, an integration of these algorithms and constraints may lead us to computational methods of approximating error covariance with improved accuracy.

Acknowledgment

We thank Professor Qing Zhang, University of Georgia, and Professor Jiangang Ying, Fudan University, for their comments and suggestions on several questions related to error covariance and stochastic processes.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported in part by U.S. Naval Research Laboratory - Monterey, CA.

References

- Bannister, R. N. 2008a. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Q. J. R. Meteorol. Soc.* **134**, 1951–1970. doi:10.1002/qj.339
- Bannister, R. N. 2008b. A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. *Q. J. R. Meteorol. Soc.* **134**, 1971–1996. doi:10.1002/qj.340
- Deleanu, D. and Dumitrache, C. L. 2019. Explicit finite difference schemes for solving the one-dimensional tsunami wave propagation equations. *J. Phys. Conf. Ser.* **1297**, 012005. doi:10.1088/1742-6596/1297/1/012005
- Gaspari, G. and Cohn, S. E. 1999. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125**, 723–757. doi:10.1002/qj.49712555417
- Gelb, A. and The Technical Staff. 1974. *Applied Optimal Estimation*. MIT Press, Cambridge, Massachusetts, USA.
- Houtekamer, P. L. and Zhang, F. 2016. Review of the ensemble Kalman filter for atmospheric data assimilation. *Month. Wea. Rev.* **144**, 4489–4532. doi:10.1175/MWR-D-15-0440.1
- Iserles, A. 2000. How large is the exponential of a banded matrix. *New Zealand J. Math.* **29**, 177–192.
- Julier, S. and Uhlmann, J. 2004. Unscented filtering and nonlinear estimation. *Proc. IEEE* **92**, 401–422. doi:10.1109/JPROC.2003.823141
- Julier, S., Uhlmann, J. and Durrant-Whyte, H. F. 2000. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Trans. Automat. Contr.* **45**, 477–482. doi:10.1109/9.847726
- Kailath, T. 1980. *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall.
- Kang, W. and Xu, L. 2009. Computational analysis of control systems using dynamic optimization. arXiv: 0906.0215v2, <https://arxiv.org/abs/0906.0215>.
- Kang, W. and Xu, L. 2012. Optimal placement of mobile sensors for data assimilations. *Tellus A* **64**, 17133. doi:10.3402/tellusa.v64i0.17133

- Krener, A. J. and Ide, K. 2009. Measures of unobservability. In 48th IEEE Conference on Decision and Control, Shanghai, China, 6401–6406.
- LeVeque, R. J. 1992. *Numerical Methods for Conservation Laws*. Birkhäuser Verlag, Basel, Boston, Berlin.
- Ménétrier, B., Montmerle, T., Michel, Y. and Berre, L. 2015. Linear filtering of sample covariances for ensemble-based data assimilation. Part I: optimality criteria and application to variance filtering and covariance localization. *Month. Wea. Rev.* **143**, 1622–1643. doi:10.1175/MWR-D-14-00157.1
- Yoshimura, R., Yakeno, A., Misaka, T. and Obayashi, S. 2020. Application of observability Gramian to targeted observation in wrf data assimilation. *Tellus A* **72**, 1–11. doi:10.1080/16000870.2019.1697602

Appendix: the proof of Theorem 1

In the following, we prove the following statement that implies the result in Theorem 1. Suppose that A , the matrix in (1), is banded with a bandwidth $s \geq 1$. Let ρ be an upper bound of the entries,

$$\rho = \max_{0 \leq i, j \leq n} \{|A_{i,j}|\}.$$

Suppose Q is a banded matrix. Let $P(t)$ be the error covariance of the Kalman-Bucy filter (2) in the time interval $[0, t]$. Then

$$|G_{ij}^C(t)| \leq \bar{G} c(\rho e^2 t, (|j-i|-L)/s), \text{ if } |j-i| > L_1 \quad (41)$$

for some constant $\bar{G} > 0$ and some integers $L, L_1 \leq n$. The following theorem from Iserles (2000) is fundamental to the proof of Theorem 1.

Theorem 2. (Iserles, 2000) Let $E = e^A$, where $A \in \mathbb{R}^{n \times n}$ is a banded matrix of bandwidth $s \geq 1$. Let

$$\rho = \max_{0 \leq i, j \leq n} \{|A_{i,j}|\},$$

then

$$|E_{i,j}| \leq c(\rho, |i-j|/s) \left(e^{|i-j|/s} - \sum_{m=0}^{|i-j|-1} \frac{(|i-j|/s)^m}{m!} \right), \quad |i-j| > L, \quad (42)$$

for some integer $L > 0$.

Because

$$\sum_{m=0}^N \frac{(|i-j|/s)^m}{m!}$$

is a sequence that is increasing and approaching $e^{|i-j|/s}$ when $N \rightarrow \infty$, (42) implies the following inequality

$$|E_{i,j}| \leq c(\rho e, |i-j|/s), \quad |i-j| > L. \quad (43)$$

Remark 1. For problems with large dimensions, the proof in Iserles (2000) assumes that $\rho \ll n$. This assumption

leads to an estimation in which $L \ll n$. However, there is no computational formula of estimating L .

Lemma 2. Let E and Z be $n \times n$ matrixes satisfying (43) for some integers L_E and L_Z , respectively. Let

$$L = \max\{L_E, L_Z, spe\} \quad (44a)$$

$$\beta = \max\{E_{ij}, Z_{ij} | 1 \leq i, j \leq n\}. \quad (44b)$$

For all $1 \leq i, j \leq n$ satisfying $|i-j| > 2L$, we have

$$|(EZ)_{ij}| \leq \frac{2}{1 - (spe/L)^{1/s}} c(\rho e, |j-i|/s) \quad (45a)$$

$$+ 2\beta(L+1)c(\rho e, |j-i|/s) \quad (45b)$$

$$+ 2\beta L c(\rho e, (|j-i|-L)/s) \quad (45c)$$

$$+ (|j-i|-2L-1)c(\rho e^2, |j-i|/s) \quad (45d)$$

Proof. Without loss of generality, we assume $i < j$. More specifically, the index $1 \leq k \leq n$ has the following partition $1 < i-L < i < i+L \leq j-L < j < j+L < n$.

Using this partition, we group the terms in $(EZ)_{ij}$

$$(EZ)_{ij} = \sum_{k=1}^{i-L-1} E_{ik} Z_{kj} + \sum_{k=i-L}^i E_{ik} Z_{kj} + \sum_{k=i+1}^{i+L} E_{ik} Z_{kj} + \sum_{k=i+L+1}^{j-L-1} E_{ik} Z_{kj} \quad (46a)$$

$$+ \sum_{k=j+L+1}^n E_{ik} Z_{kj} + \sum_{k=j}^{j+L} E_{ik} Z_{kj} + \sum_{k=j-L}^{j-1} E_{ik} Z_{kj} \quad (46b)$$

In the following, we derive the upper bounds for every summation on the right-hand side of the equation in (46). The four summations in the top row represent four different cases. The summations in the second row are aligned with the first row in such a way that each summation has the same upper bound as that of the summation on top.

Case I, $1 \leq k < i-L$. It is obvious that $|j-k|/s > (j-i)/s$

$$\text{From (16) in Lemma 1, we have} \quad \left| \sum_{k=1}^{i-L-1} E_{ik} Z_{kj} \right| \leq \sum_{k=1}^{i-L-1} c(\rho e, |i-k|/s) c(\rho e, |j-k|/s) \quad (47a)$$

$$= \sum_{k=1}^{i-L-1} \frac{(\rho e)^{|i-k|/s}}{(|i-k|/s)^{|i-k|/s}} c(\rho e, |j-k|/s) \quad (47b)$$

$$\leq \sum_{k=1}^{i-L-1} \frac{(\rho e)^{|i-k|/s}}{(L/s)^{|i-k|/s}} c(\rho e, |j-i|/s), \quad (47c)$$

$$\begin{aligned} & \text{(because } |i-k| \geq L \text{ because } |j-i| < |j-k|) \\ & \leq \frac{1}{1 - (spe/L)^{1/s}} c(\rho e, |j-i|/s) \end{aligned} \quad (47d)$$

It equals one half of the first term on the right-hand side of (45). Similarly, one can prove the same inequality for $j + L + 1 \leq k \leq n$.

Case II, $i - L \leq k \leq i$. In this case we have

$$\left| \sum_{k=i-L}^i E_{ik} Z_{kj} \right| \leq \beta \sum_{k=i-L}^i c(\rho e, |j-k|/s) \quad (48a)$$

$$\leq \beta \sum_{k=i-L}^i c(\rho e, |j-i|/s) \quad (48b)$$

$$= \beta(L+1)c(\rho e, |j-i|/s) \quad (48c)$$

It equals one half of the second term on the right-hand side of (45). Similarly, one can prove the same inequality for $j \leq k \leq j + L$.

Case III, $i + 1 < k \leq i + L$. Because $L \leq j - i - L \leq j - k$, then

$$\left| \sum_{k=i+1}^{i+L} E_{ik} Z_{kj} \right| \leq \beta \sum_{k=i+1}^{i+L} c(\rho e, |j-k|/s) \quad (49a)$$

$$\leq \beta \sum_{k=i+1}^{i+L} c(\rho e, (j-i-L)/s) \quad (49b)$$

$$(\text{because } L \leq j - i - L \leq j - k)$$

$$= \beta L c(\rho e, (j-i-L)/s), \quad (49c)$$

It equals one half of the third term on the right-hand side of (45). Similarly, one can prove the same inequality for $j - L \leq k \leq j - 1$.

Case IV, $i + L < k < j - L$. If $|j-i| = 2L+1$, this is an empty index set. The coefficient in the corresponding term equals zero. Let's assume that $|j-i| > 2L+1$. It is known that $(1 + \alpha/x)^x$, $x > 0$, is an increasing function of x . It approaches e^α when $x \rightarrow \frac{\alpha}{x} \infty$, i.e.

$$\left(1 + \frac{\alpha}{x}\right) \uparrow e. \quad (50)$$

Consider

$$\left| \sum_{k=i+L+1}^{j-L-1} E_{ik} Z_{kj} \right| \leq \sum_{k=i+L+1}^{j-L-1} \frac{(\rho e)^{(k-i)/s}}{((k-i)/s)^{(k-i)/s}} \frac{(\rho e)^{(j-k)/s}}{((j-k)/s)^{(j-k)/s}} \quad (51a)$$

$$= \sum_{k=i+L+1}^{j-L-1} \frac{(\rho e)^{(k-i)/s} \left(1 + \frac{j-k}{k-i}\right)^{(k-i)/s}}{\left(\frac{k-i}{s}\right)^{(k-i)/s} \left(1 + \frac{j-k}{k-i}\right)^{(k-i)/s}} \quad (51b)$$

$$\frac{(\rho e)^{(j-k)/s} \left(1 + \frac{k-i}{j-k}\right)^{(j-k)/s}}{\left(\frac{j-k}{s}\right)^{(j-k)/s} \left(1 + \frac{k-i}{j-k}\right)^{(j-k)/s}} \quad (51c)$$

$$= \sum_{k=i+L+1}^{j-L-1} \frac{(\rho e)^{(j-i)/s} \left(1 + \frac{j-k}{k-i}\right)^{(k-i)/s} \left(1 + \frac{k-i}{j-k}\right)^{(j-k)/s}}{((j-i)/s)^{(k-i)/s} ((j-i)/s)^{(j-k)/s}} \quad (51d)$$

Applying (50), we have

$$\left| \sum_{k=i+L+1}^{j-L-1} E_{ik} Z_{kj} \right| \leq \sum_{k=i+L+1}^{j-L-1} \frac{(\rho e)^{(j-i)/s} e^{(j-k)/s} e^{(k-i)/s}}{((j-i)/s)^{(j-i)/s}} \quad (52a)$$

$$\leq \sum_{k=i+L+1}^{j-L-1} \frac{(\rho e^2)^{(j-i)/s}}{((j-i)/s)^{(j-i)/s}} \quad (52b)$$

$$\leq (j-i-2L-1)c(\rho e^2, (j-i)/s) \quad (52c)$$

It equals the 4th term on the right-hand side of (45). $\square$

The Proof of Theorem 1. In the following, we prove (41).

We first assume that $Q=I$ is identity. Let $E(t)$ denote the exponential matrix

$$E(t) = e^{At}$$

According to Theorem 2 and the inequality (43),

$$|E(t)_{ij}| \leq c(\rho e t, |i-j|/s), \quad \text{if } |i-j| > L_1 \quad (53)$$

for some $L_1 > 0$. Obviously, the same inequality holds true for the matrix $e^{A^T t}$. From Lemma 2,

$$|(e^{At} Q e^{A^T t})_{ij}| \leq \frac{2}{1 - (s\rho e/L)^{1/s}} c(\rho e t, |j-i|/s) \quad (54a)$$

$$+ 2\beta(L+1)c(\rho e t, |j-i|/s) \quad (54b)$$

$$+ 2\beta L c(\rho e t, (|j-i|-L)/s) \quad (54c)$$

$$+ (|j-i|-2L-2)c(\rho e^2 t, |j-i|/s),$$

$$\text{if } |j-i| > 2L$$

$$(54d)$$

where

$$L = \max\{s\rho e T, L_1\}, \quad (55a)$$

$$\beta = \max\{|(e^{At})_{ij}| \mid 1 \leq i, j \leq n, t \in [0, T]\} \quad (55b)$$

Because

$$c(\alpha, x)$$

is an increasing function of α and a decreasing function of x , (54) implies

$$|(e^{At} Q e^{A^T t})_{ij}| \leq \bar{G}_1 c(\rho e^2 T, (|j-i|-L)/s), \quad \text{if } |j-i| > 2L \quad (56a)$$

for some $\bar{G}_1 > 0$. Substituting (56) into the controllability gramian,

$$|G_{ij}| \leq \int_0^T \bar{G}_1 c(\rho e^2 T, (|j-i|-L)/s) d\xi \quad (57a)$$

$$= \bar{G}_1 T c(\rho e^2 T, (|j-i|-L)/s) \quad (57b)$$

This is equivalent to (41). If Q is banded with bandwidth s_Q , then Q equals the sum of $2s_Q + 1$ diagonal matrices along the main diagonal and s_Q subdiagonals. The subdiagonals in Q do not change the expression in the inequality (56) except that it increases $\bar{G}$ and L . $\square$