An Extended Kalman Filter with Updated Noise Covariance for Parameter Estimation in Chemical Reaction Networks

Suryasnata Dash[1], Appana Sai Sasi Kumar[2] and Abhishek Dey[3]

***Abstract*— Parameter estimation in chemical reaction networks is a challenging task due to its inherent nonlinearity and stochasticity. Extended Kalman Filters (EKF) have been widely used for this purpose. However, the process noise covariance in the Kalman filter algorithm is hard to determine and the effect of reduced order modelling on estimation is generally unknown. Here, we implement a Continuous Discrete-EKF (CD-EKF) with process noise covariance updated based on Chemical Langevin Equation (CLE). Further, we analyze the performance of the proposed filter, both using full and reduced order models. We find that the filter performance is better compared to fixed choices of noise covariance based on whiteness tests and the filter achieves a balance between mean squared estimation error and parameter convergence time.**

I. INTRODUCTION

Biological systems are inherently noisy and mostly nonlinear in nature. Biological functions are generally described through reactions, involving different species, known as Chemical Reaction Networks (CRN). CRNs are widely represented in a deterministic approach with Ordinary Differential Equations (ODE). It is often desired to obtain an approximate reduced order model due to large number of species involved in CRNs. The reduced model can be obtained through time-scale separation methods like Quasi-Equilibrium Approximation (QEA) and separation methods such as Quasi-Steady-State Approximation (QSSA) [1]. Although ODE models are useful, a more accurate stochastic representation is necessary to incorporate the noise present in CRNs. Chemical Master Equations (CME) are used for this purpose which represents temporal evolution of species concentration probabilities at each time point. [2]. Various approximation of CMEs are used in stochastic analysis of CRNs such as Chemical Langevin Equations (CLE) and system size expansion [3].

In CRN models, all parameters may not be known, and estimation techniques are adopted to determine them. Due to nonlinear nature of CRNs, Extended Kalman Filter (EKF) has been used for parameter estimation [4]. A comparative study regarding estimation of states and parameters for a *Cad* system of *E. coli* has been done with EKF, Unscented Kalman filter, Particle filter [5]. Hybrid Extended Kalman filter (HEKF) has been used for parameter estimation in heat shock response of *E. coli* [6]. Parameters of biological functions like signal transduction, mitogen-activated protein kinase signalling pathway were estimated using Bayesian methods [7]. EKF is also applied to estimate the posterior distribution of low order moments of states in bacterial two-component regulatory systems obtained by Monte Carlo simulations [8]. These studies provide important groundwork for parameter estimation in CRNs.

Modelling of the noise in stochastic system significantly affects parameter estimation and having prior knowledge is helpful [9]. Specifically, in case of parameter estimation using EKF, the choice of process noise covariance is important [10]. Typically, the process noise covariance is assumed to be constant which is decided based on trial and error. Although there is abundant literature that uses Adaptive Kalman filter (AKF) techniques to estimate the noise covariances, our approach is to determine the noise covariances for CRNs from the first principles model such as CLE. A previous study considered an estimate-updated process noise covariance for biological systems based on Langevin formalism [11]. However, the covariance matrix was only restricted to a diagonal one. Further, the effect of reduced order CRN models on EKF performance while using this updated noise covariance was not considered.

To address these questions, we propose a Continuous Discrete-Extended Kalman Filter (CD-EKF) with an updated process noise covariance based on CLE models of CRNs. We take examples of two widely occurring biological systems and estimate unknown parameters from data generated using stochastic simulation and incorporate the updated process noise covariance. We find that the proposed filter performance is better compared to fixed choices of process noise covariance both for the full order and reduced order models. Further, the proposed filter achieves a balance between the Normalized Root Mean Square Error (NRMSE) and parameter convergence time. These results can provide better insight in designing filters to estimate parameters in stochastic systems.

The next section of the paper contains basic definitions and tools required in the process, while Section III contains the mathematical models involved in estimation of parameters, Section IV discusses the results and finally, Section V gives an overview of the analysis and concludes the investigation.

II. DEFINITIONS

A. Deterministic CRN model

A CRN model, containing species $X = [X_1, X_2...X_N]$ and reactions $r = [r_1, r_2...r_\mu]$ with unknown model parameter set θ, can have dynamics represented in deterministic

[1,3]Suryasnata Dash and Abhishek Dey are with Department of Electrical Engineering, National Institute of Technology Rourkela, Sector 1, Rourkela, 769008, India 523ee1001@nitrkl.ac.in, deyab@nitrkl.ac.in

[2]Appana Sai Sasi Kumar is with PVInsight Pvt. Ltd., Hitech city, Hyderabad, Telangana, 500081, India sasikumar.appana@pvinsightinc.com

form for species X_i as,

$$\frac{dX_i}{dt} = \sum_{j=1}^{\mu} a_j(X, \alpha) v_{ij}. \tag{1}$$

Here, a_j represents the propensity of j^{th} reaction, as function of X and $\alpha \subset \theta$, v_{ij} represents the stoichiometric coefficient. The measurement equation is represented in discrete time steps as,

$$y_k = H_k X_k. \tag{2}$$

B. Chemical Langevin Equation

Stochastic representation of the system in (1) can have various approximations and one of the most popular one is CLE [12]. The approximation gives rise to a mean term and a noise term extending (1) to,

$$\frac{dX_i}{dt} = f(X, \alpha) + G(X, \alpha). \tag{3}$$

Here, $f(X, \alpha) = \sum_{j=1}^{\mu} a_j(X, \alpha) v_{ij}$ and $G(X, \alpha) = \sum_{j=1}^{\mu} v_{ij} \sqrt{a_j(X, \alpha)} \Gamma_j$, where Γ_j is white Gaussian noise corresponding to j^{th} reaction.

C. Continuous Discrete-Extended Kalman Filter

CD-EKF involves the system model given by (3), while the output equation given by (2) [13]. The principles of EKF are applied on the model. The nonlinear system dynamics are linearized around neighbourhood of the estimates,

$$\begin{aligned} \frac{dx}{dt} &= Fx + L\Gamma \\ y_k &= H_k x_k + v_k. \end{aligned} \tag{4}$$

Γ and v represent the zero mean process and measurement Gaussian noise with covariance matrices Q and R respectively. F and L are the Jacobian matrices of the model in (3) with respect to states and process noise. The priori estimates and error covariance are calculated integrating equations (5) from t_{k-1} to t_k,

$$\begin{aligned} \frac{d\hat{x}_k^-}{dt} &= f(\hat{x}_{k-1}^+) \\ \frac{dP_k^-}{dt} &= F_{k-1} P_{k-1}^+ + P_{k-1}^+ F_{k-1}^T + LQL^T. \end{aligned} \tag{5}$$

The posterior estimates and error covariance are calculated at t_k as,

$$\begin{aligned} \hat{x}_k^+ &= \hat{x}_k^- + K(y_k - H_k \hat{x}_k^-) \\ P_k^+ &= (I - KH_k) P_k^-. \end{aligned} \tag{6}$$

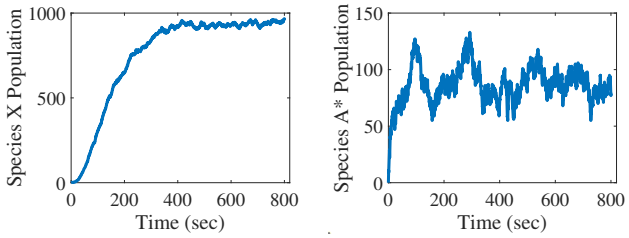
K is the Kalman gain of the system calculated as,

$$K = P_k^- H^T (H_k P_k^- H_k^T + R)^{-1}. \tag{7}$$

III. MATHEMATICAL MODELLING

To assess the performance of proposed filter, we have taken two representative systems of CRN - 1) Gene expression system 2) Biomolecular covalent modification system:

Data generation: To estimate the unknown states and parameters of the system we have used data generated by Stochastic simulation algorithm (SSA) [2]. The algorithm generates the dynamics of the reactions with noise involved and represents an exact behaviour of the reactions. The simulated protein (X) and activated substrate (A^*) population are used as data in Example 1 and 2 respectively for estimation and error analysis in CD-EKF. The simulations for the examples were run for 800 seconds (Fig. 1). For both the systems, only one variable is observed which is generally true in case of biomolecular systems. Due to this, estimating more than one unknown parameter using CD-EKF has observability issues. For this reason we have considered the degradation rate constant of protein (d_X) in Example 1 and the production rate of activated substrate (k_1) in Example 2 to be unknown. The measurement noise is uniformly distributed in range $[-5, 5]$ and 50 datasets were generated for each example.



(a) (b)

Fig. 1: A single dataset generated for a) Gene expression model, b)Biomolecular covalent modification model by stochastic simulation algorithm. The parameter regime for a) $P_{tot} = 100$, $R_{tot} = 100$, $G = 10$, $ktx = ktl = 0.1$, $kup = kur = 0.05$, $kbp = kbr = 0.02$, $d_T = 0.01$, $d_X = 0.01$, b)$E_{1tot} = E_{2tot} = 10$, $A_{tot} = 200$, $\alpha_1 = \alpha_2 = 0.1$, $d_1 = d_2 = 1$, $k_1 = k_2 = 1$.

Next, we describe both full order and reduced order stochastic model for these two systems.

A. Example 1 - Gene Expression

Gene expression model consists of two phases of reactions. The first phase occurs when a DNA strand (G) forms a complex (C_1) with help of enzyme RNA polymerase (P_o) and the complex produces the mRNA transcript (T). The second phase reaction is when the mRNA transcript forms another complex (C_2) with Ribosome (R_i) and the C_2 produces protein (X). It is defined by the following chemical reactions [1].

$$\mathrm{G} + P_o \underset{\text{kup}}{\overset{\text{kbp}}{\rightleftharpoons}} \mathrm{C}_1$$

$$\mathrm{C}_1 \xrightarrow{\text{ktx}} \mathrm{G} + P_o + \mathrm{T}$$

$$\mathrm{T} + R_i \underset{\mathrm{kur}}{\overset{\mathrm{kbr}}{\rightleftharpoons}} \mathrm{C}_2$$
$$\mathrm{C}_2 \xrightarrow{\mathrm{ktl}} \mathrm{T} + R_i + \mathrm{X}$$
$$\mathrm{T} \xrightarrow{\mathrm{d_T}} \emptyset$$
$$\mathrm{X} \xrightarrow{\mathrm{d_X}} \emptyset\,.$$

The mRNA polymerase and ribosomes population are conserved in this example as $P_{tot} = P_o + C_1$, $R_{tot} = R_i + C_2$.

The CLE of the process are represented as following stochastic differential equations,

$$\begin{aligned}
\frac{dP_o}{dt} &= (k_{up} + k_{tx})(P_{tot} - P_o) - k_{bp} G P_o - \sqrt{k_{bp} G P_o}\Gamma_1 \\
&\quad + \sqrt{k_{up}(P_{tot} - P_o)}\Gamma_2 + \sqrt{k_{tx}(P_{tot} - P_o)}\Gamma_3 \\
\frac{dT}{dt} &= k_{tx}(P_{tot} - P_o) + (k_{tl} + k_{ur})(R_{tot} - R_i) - k_{br} T R_i \\
&\quad - d_T T + \sqrt{k_{tx}(P_{tot} - P_o)}\Gamma_3 - \sqrt{k_{br} T R_i}\Gamma_4 \\
&\quad + \sqrt{k_{ur}(R_{tot} - R_i)}\Gamma_5 + \sqrt{k_{tl}(R_{tot} - R_i)}\Gamma_6 - \sqrt{d_T T}\Gamma_7 \\
\frac{dR_i}{dt} &= (k_{ur} + k_{tl})(R_{tot} - R_i) - k_{br} T R_i - \sqrt{k_{br} T R_i}\Gamma_4 \\
&\quad + \sqrt{k_{ur}(R_{tot} - R_i)}\Gamma_5 + \sqrt{k_{tl}(R_{tot} - R_i)}\Gamma_6 \\
\frac{dX}{dt} &= k_{tl}(R_{tot} - R_i) - d_X X + \sqrt{k_{tl}(R_{tot} - R_i)}\Gamma_6 \\
&\quad - \sqrt{d_X X}\Gamma_8. \qquad (8)
\end{aligned}$$

Equation (8) represents the full order model of the process. Under QSSA, the complexes are assumed to reach the steady state faster than other molecules resulting to, $\frac{dC_1}{dt} = 0$, $\frac{dC_2}{dt} = 0$.

Through these assumptions, hill function propensity based reduced order deterministic model is formed for the system [1]. The stochastic representation of the reduced order deterministic model can be,

$$\begin{aligned}
\frac{d\hat{T}}{dt} &= k_{tx} P_{tot}\left(\frac{G}{K_1 + G}\right) - d_T \hat{T} + \sqrt{k_{tx} P_{tot}\left(\frac{G}{K_1 + G}\right)}\Gamma_1 \\
&\quad - \sqrt{d_T \hat{T}}\Gamma_2 \\
\frac{d\hat{X}}{dt} &= k_{tl} R_{tot}\left(\frac{\hat{T}}{K_0 + \hat{T}}\right) - d_X \hat{X} + \sqrt{k_{tl} R_{tot}\left(\frac{\hat{T}}{K_0 + \hat{T}}\right)}\Gamma_3 \\
&\quad - \sqrt{d_X \hat{X}}\Gamma_4. \qquad (9)
\end{aligned}$$

where, $K_1 = \dfrac{k_{tx} + k_{up}}{k_{bp}}$, $K_0 = \dfrac{k_{tl} + k_{ur}}{k_{br}}$.

B. Example 2 - Biomolecular Covalent Modification System

Biomolecular covalent modification system, also known as Goldbeter-Koshland process contains two enzymatic reactions, transitioning a molecule to its excited state and vice-versa, where the change of reactants are affected by rate constants [14]. The change in parameters determine the steepness of the input-output (product concentration vs. rate of reaction) curve in the system [15]. Most common example of this system includes phosphorylation and dephosphorylation. The chemical reactions are represented as,

$$\mathrm{A} + \mathrm{E}_1 \underset{\mathrm{d_1}}{\overset{\alpha_1}{\rightleftharpoons}} \mathrm{C}_1$$
$$\mathrm{C}_1 \xrightarrow{\mathrm{k_1}} A^* + \mathrm{E}_1$$
$$A^* + \mathrm{E}_2 \underset{\mathrm{d_2}}{\overset{\alpha_2}{\rightleftharpoons}} \mathrm{C}_2$$
$$\mathrm{C}_2 \xrightarrow{\mathrm{k_2}} \mathrm{A} + \mathrm{E}_2.$$

The total enzyme and substrate population are conserved in the reactions as $E_{1tot} = E_1 + C_1$, $E_{2tot} = E_2 + C_2$, $A_{tot} = A + A^* + C_1 + C_2$.

The CLE of the system is represented as following stochastic differential equations,

$$\begin{aligned}
\frac{dA}{dt} &= -\alpha_1 A(E_{1tot} - C_1) + d_1 C_1 \\
&\quad + k_2(A_{tot} - A - A^* - C_1) - \sqrt{\alpha_1 A(E_{1tot} - C_1)}\Gamma_1 \\
&\quad + \sqrt{d_1 C_1}\Gamma_2 + \sqrt{k_2(A_{tot} - A - A^* - C_1)}\Gamma_6 \\
\frac{dC_1}{dt} &= \alpha_1 A(E_{1tot} - C_1) - d_1 C_1 - k_1 C_1 \\
&\quad + \sqrt{\alpha_1 A(E_{1tot} - C_1)}\Gamma_1 - \sqrt{d_1 C_1}\Gamma_2 - \sqrt{k_1 C_1}\Gamma_3 \\
\frac{dA^*}{dt} &= -\alpha_2 A^*(E_{2tot} - A_{tot} + A + A^* + C_1) + k_1 C_1 \\
&\quad + d_2(A_{tot} - A - A^* - C_1) + \sqrt{k_1 C_1}\Gamma_3 \\
&\quad - \sqrt{\alpha_2 A^*(E_{2tot} - A_{tot} + A + A^* + C_1)}\Gamma_4 \\
&\quad + \sqrt{d_2(A_{tot} - A - A^* - C_1)}\Gamma_5. \qquad (10)
\end{aligned}$$

For reduced order model, the total enzyme population are considered to be negligible as compared to substrate ($E_{1tot} = E_{2tot} << A_{tot}$) [14] and the conservation law is reduced to, $A_{tot} = A + A^*$. The complexes are assumed to reach the steady state faster than substrate, reducing the overall system. Hence, the CLE representation of reduced order model with QSSA is represented as,

$$\begin{aligned}
\frac{dA^*}{dt} &= k_1 E_{1tot}\frac{A_{tot} - A^*}{A_{tot} - A^* + K_{m1}} - k_2 E_{2tot}\frac{A^*}{A^* + K_{m2}} \\
&+ \sqrt{k_1 E_{1tot}\frac{A_{tot} - A^*}{A_{tot} - A^* + K_{m1}}}\Gamma_1 - \sqrt{k_2 E_{2tot}\frac{A^*}{A^* + K_{m2}}}\Gamma_2, \qquad (11)
\end{aligned}$$

where, $K_{m1} = \dfrac{k_1 + d_1}{\alpha_1}$, $K_{m2} = \dfrac{k_2 + d_2}{\alpha_2}$.

Proposed filter: We use CLE representation of the reaction networks in CD-EKF. Estimation is done through augmenting unknown parameters as states, $\frac{d\theta}{dt} = \eta(t)$, where $\eta(t)$ is a zero mean white Gaussian noise with small variance ζ [9]. We define $\hat{z}_{k-1} = [\hat{X}_{k-1}, \hat{\theta}_{k-1}]^T$ with $\hat{z}_{k-1} \in R^n$ and overall covariance matrix is described as,

$$\hat{Q}(\hat{z}_{k-1}) = \left[\begin{array}{c|c} L(\hat{z}_{k-1}) Q L^T(\hat{z}_{k-1}) & 0 \\ \hline 0 & \zeta \end{array}\right].$$

$L(\hat{z}_{k-1}) Q L^T(\hat{z}_{k-1})$ is modified process-noise covariance of the system with Q assumed as identity matrix. It is assumed that system state noise and unknown parameter noise are uncorrelated. $L(\hat{z}_{k-1})$ for system in (8) and (10) are given as $L_G(\hat{z}_{k-1})$ and $L_B(\hat{z}_{k-1})$ in (12) and (13) respectively. Similarly, $L(\hat{z}_{k-1})$ can be constructed for the CLE

$$L_G(\hat{z}_{k-1}) = \begin{bmatrix} -\sqrt{k_{bp}GP_o} & \sqrt{k_{up}(P_{tot}-P_o)} & \sqrt{k_{tx}(P_{tot}-P_o)} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{k_{tx}(P_{tot}-P_o)} & -\sqrt{k_{br}TR_i} & 0 & \sqrt{k_{tl}(R_{tot}-R_i)} & -\sqrt{d_T T} & 0 \\ 0 & 0 & 0 & -\sqrt{k_{br}TR_i} & \sqrt{k_{ur}(R_{tot}-R_i)} & \sqrt{k_{tl}(R_{tot}-R_i)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{k_{tl}(R_{tot}-R_i)} & 0 & -\sqrt{d_X X} \end{bmatrix}, \tag{12}$$

$$L_B(\hat{z}_{k-1}) = \begin{bmatrix} -\sqrt{\alpha_1 A(E_{1tot}-C_1)} & \sqrt{d_1 C_1} & 0 & 0 & 0 & \sqrt{k_2(A_{tot}-A-A^*-C_1)} \\ \sqrt{\alpha_1 A(E_{1tot}-C_1)} & -\sqrt{d_1 C_1} & \sqrt{k_1 C_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{k_1 C_1} & -\sqrt{\alpha_2 A^*(E_{2tot}-A_{tot}+A+A^*+C_1)} & \sqrt{d_2(A_{tot}-A-A^*-C_1)} \end{bmatrix}. \tag{13}$$

representations of reduced order models of both reaction networks. $L(\hat{z}_{k-1})QL^T(\hat{z}_{k-1})$ is a non-diagonal matrix, and it is considered so as per the definition of covariance matrix in [13]. We update $L(\hat{z}_{k-1})$ in each iteration of CD-EKF algorithm based on previous estimates of unknown states and parameter to numerically integrate (5). The measurement noise covariance matrix R is considered constant and known.

IV. Results and Discussion

Next, we analysed the performance of the proposed filter with different fixed choices of process noise covariance Q and calculated couple of statistics. First, we find the parameter convergence time using notion of convergence in mean. To calculate the convergence time for the estimated parameter, first we find an array of parameter values in consecutive data points for which the mean of the array,

$$\theta_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\theta_i, \tag{14}$$

is within $\pm 5\%$ error band of true value. Then, we define parameter convergence time as the time corresponding to the first element of this array. Second, the NRMSE is calculated as,

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - H\hat{z_k})^2}}{\frac{1}{N}\sum_{j=1}^{N} y_k}. \tag{15}$$

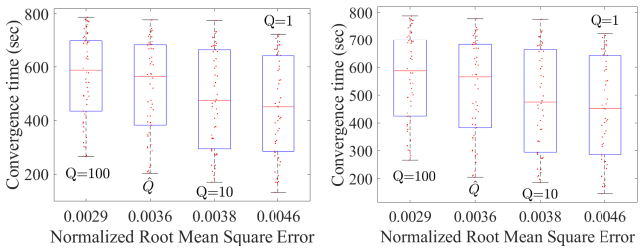
A. Trade off between estimation error and parameter convergence time

1) Example 1 - Gene expression model: In case of gene expression model, we assumed that the protein degradation parameter (d_X) is unknown and other parameters are known. The statistics obtained for 50 datasets in case of full order and reduced order gene expression system are tabulated in Table I. For different Q values, the mean convergence time increases with increasing Q while the mean NRMSE increases with decreasing Q in both full order and reduced order model (Fig. 2). In both models, the proposed filter with $\hat{Q}$ matrix achieves a balance between mean convergence time and mean NRMSE.

2) Example 2 - Biomolecular covalent modification system: In case of biomolecular covalent modification system, we assumed that the parameter k_1 is unknown and rest of the parameters are known. In case of full order model, we observe similar trade off between convergence time and NRMSE (Table II, Fig. 3a). In the reduced order model, two cases are taken into consideration since the approximation

TABLE I: Convergence time and NRMSE of full and reduced order gene expression model

	Process Noise Covariance (Q)	Convergence Time (seconds)		Normalized Root Mean Square Error	
		Mean	Standard deviation	Mean	Standard deviation
Full order	$1 \times I_{n\times n}$	452.928	183.496	0.0046	8.696×10^{-5}
	$10 \times I_{n\times n}$	480.410	186.019	0.0038	6.666×10^{-5}
	$\hat{Q}$	520.853	174.220	0.0036	6.303×10^{-5}
	$100 \times I_{n\times n}$	567.725	162.241	0.0029	5.085×10^{-5}
Reduced order	$1 \times I_{n\times n}$	452.746	183.983	0.0046	8.958×10^{-5}
	$10 \times I_{n\times n}$	481.484	184.441	0.0038	6.975×10^{-5}
	$\hat{Q}$	521.547	173.327	0.0036	6.071×10^{-5}
	$100 \times I_{n\times n}$	570.368	160.487	0.0029	4.863×10^{-5}



(a) (b)

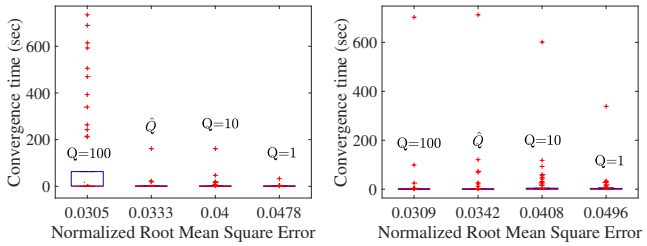
Fig. 2: Boxplot of convergence time along with the data distribution grouped by mean NRMSE for a) full order b) reduced order gene expression model.

gives a lumped parameter which is a function of the unknown parameter k_1. i) The Michaelis-Menten constant K_{m1} is constant as per the convention. ii) K_{m1} is a function of estimated k_1. The observations due to different Q values are shown with case i) on the top and case ii) recorded on bottom. A less pronounced version of the pattern in case of full order model is observed in reduced order case (Fig. 3b).The convergence time has higher variance for all the choices of Q, possibly owing to larger uncertainty in model.

The elements of $\hat{Q}$ matrix used in both examples have a transient to steady state response in mean. The updation at each instant for the elements of $\hat{Q}$ matrix ensures a

TABLE II: Convergence time and NRMSE of full and reduced order biomolecular covalent modification system

	Process Noise Covariance (Q)	Convergence Time (seconds)		Normalized Root Mean Square Error	
		Mean	Standard deviation	Mean	Standard deviation
Full order	$0.1 \times I_{n\times n}$	1.682	4.626	0.0478	0.0014
	$1 \times I_{n\times n}$	6.032	23.685	0.04	0.0011
	$\hat{Q}$	4.791	22.957	0.0333	0.0009
	$50 \times I_{n\times n}$	108.022	208.520	0.0305	0.0008
Reduced order	$0.1 \times I_{n\times n}$	10.739	48.046	0.0496	0.0015
		$K_{m1}(k_1)$			
		12.231	48.650	0.0496	0.0015
	$1 \times I_{n\times n}$	23.181	86.95	0.0408	0.0011
		$K_{m1}(k_1)$			
		20.804	86.049	0.0408	0.0011
	$\hat{Q}$	21.074	102.212	0.0342	0.0009
		$K_{m1}(k_1)$			
		20.636	101.701	0.0342	0.0009
	$50 \times I_{n\times n}$	17.26	99.920	0.0309	0.0008
		$K_{m1}(k_1)$			
		17.712	99.700	0.0309	0.0008



(a) (b)

Fig. 3: Boxplot of convergence time along with the data distribution grouped by mean NRMSE for a) full order b) reduced order biomolecular covalent modification model.

proper trade-off achieved for error and convergence time. The temporal plots of Euclidean norm are shown in Fig. 4, which can be compared to the constant Q values in Fig. 2 and 3. We also observe that in case reduced order, the steady state value of the mean Euclidean norm is less compared to the full order case possibly due to reduction in number of states.

B. Filter Optimality Tests

The performance of filter is gauged through couple of whiteness tests. These tests determine whether the noise involved in the output process is solely random or is affected through internal factors. The noise or innovation sequence being white is a necessary and sufficient condition for an optimal filter [16]. To check this we perform the autocorrelation of innovations and Ljung-Box-Pierce test.

1) Autocorrelation of innovation sequence: In the CRN system, the innovation sequence is determined as $v_k = y_k - H\hat{z}_k$. To check the whiteness of the innovation sequence, the normalized correlation,

$$\rho[l] = \frac{\sum_{i=l}^{N} v_k v_{k-i}}{\sum_{i=0}^{N} v_k v_k}, \tag{16}$$

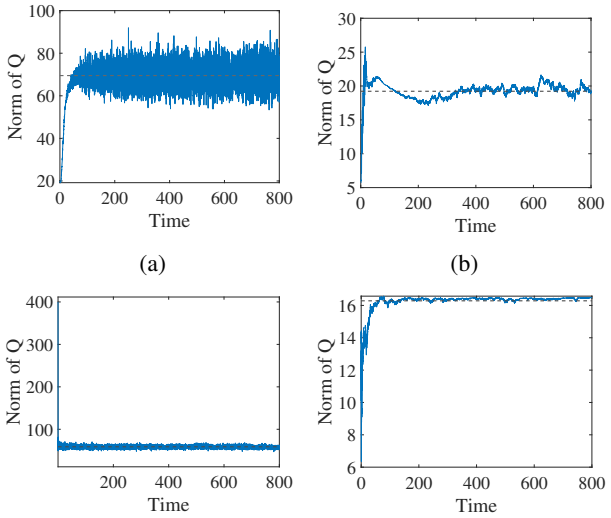


Fig. 4: Plot of Euclidean norm of $\hat{Q}$ for a) full order b) reduced order gene expression model and for c) full order d) reduced order biomolecular covalent modification model with dotted line as their mean values.

of residuals upto 100 lags are calculated. The $\rho[l]$ approaches normal distribution with 95% confidence limit if the values are within limits of $u = \pm\frac{1.96}{\sqrt{N}}$ [16]. The normalized autocorrelation for gene expression model and biomolecular covalent modification system in full order and reduced order model were calculated. In Table III, number of lags outside 95% confidence limits is minimum in case of $\hat{Q}$. In case of reduced order model of biomolecular covalent modification system, the performance of filter is lower than the gene expression system, however the choice of $\hat{Q}$ still gives a better performance compared to fixed Q values.

TABLE III: Percentage of autocorrelation values outside 95% confidence limits

		Q=1 $\times$ $I_{n\times n}$	Q=10 $\times$ $I_{n\times n}$	$\hat{Q}$	Q=100$\times$ $I_{n\times n}$
Example 1	Full order	52%	14%	0%	6%
	Reduced order	60%	17%	0%	6%
		Q=0.1$\times$ $I_{n\times n}$	Q=1 $\times$ $I_{n\times n}$	$\hat{Q}$	Q=50 $\times$ $I_{n\times n}$
Example 2	Full order	100%	68%	2%	13%
	Reduced order	100%	91%	7%	13%

2) Ljung-Box-Pierce Test: In Ljung-Box-Pierce test, total lags are considered at once and the statistic is calculated as [17],

$$\tilde{Q} = N(N+2)\sum_{i=1}^{l} \frac{\rho[i]^2}{N-i}. \tag{17}$$

where l is total lags considered, N is total number of samples, $\rho[i]$ is autocorrelation function. The statistical value should be less than chi-squared distribution of significance level α and h degrees of freedom, $\tilde{Q} < \chi_{1-\alpha,h}$.

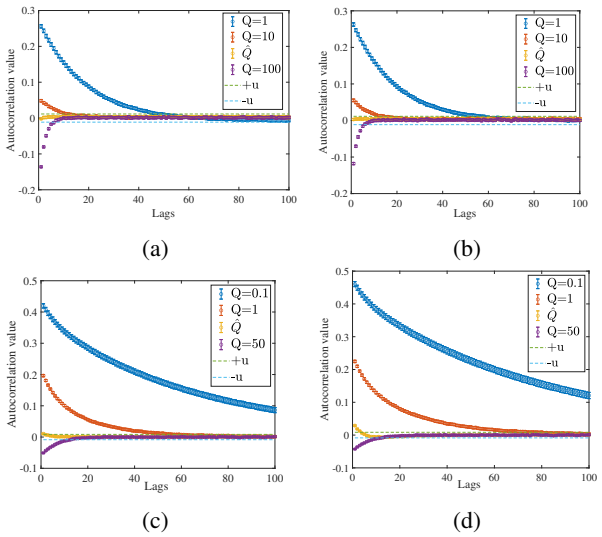


Fig. 5: Plots of autocorrelation function with respect to 100 lags. a) Full order gene expression model, b) reduced order gene expression model, c) full order biomolecular covalent modification system, d) reduced order biomolecular covalent modification system.

The significance level in this case is 5% and degrees of freedom is calculated by difference of number of unknown parameters from total lag. The number of datasets for each process which have $\tilde{Q}$ above the critical value is tabulated below (Table IV). In case of $\hat{Q}$ matrix, minimum number of datasets have $\tilde{Q}$ above the critical value. In case of gene expression model for both full order and reduced order, the number of datasets above critical value are minimum in case of $\hat{Q}$ matrix. For full order model of biomolecular covalent modification system, the number of datasets above the critical value is minimum in case of $\hat{Q}$ matrix but remains same as other Q values in case of reduced order model.

TABLE IV: Number of datasets (out of 50) with $\tilde{Q}$ above the critical value

Example 1 (l = 100)		Q = 1 × $I_{n\times n}$	Q = 10 × $I_{n\times n}$	$\hat{Q}$	Q = 100× $I_{n\times n}$
	Full order	50	50	18	50
	Reduced order	50	50	26	50
Example 2 (l = 100)		Q = 0.1 × $I_{n\times n}$	Q = 1 × $I_{n\times n}$	$\hat{Q}$	Q = 50 × $I_{n\times n}$
	Full order	50	50	11	50
	Reduced order	50	50	50	50

V. CONCLUSIONS

In this paper, we have considered two ubiquitous biochemical processes, gene expression and biomolecular covalent modification system. We have developed an EKF with updated process noise covariance based on the Langevin equation to estimate unknown parameter for these two processes. We observed that the developed filter is more optimal compared to fixed choices of noise covariance based on whiteness tests. Moreover, this filter achieves a balance between mean squared estimation error and parameter convergence time. Further, we used a reduced order stochastic model for both of this processes and we got similar optimality result as in the full order model. These results can improve our understanding in designing estimators for stochastic systems.

REFERENCES

[1] A. Pandey and R. M. Murray, "Robustness guarantees for structured model reduction of dynamical systems with applications to biomolecular models," *International Journal of Robust and Nonlinear Control*, vol. 33, no. 9, pp. 5058–5086, 2023.

[2] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[3] D. Schnoerr, G. Sanguinetti, and R. Grima, "Approximation and inference methods for stochastic biochemical kinetics—a tutorial review," *Journal of Physics A: Mathematical and Theoretical*, vol. 50, no. 9, p. 093001, 2017.

[4] Z. Wang, X. Liu, Y. Liu, J. Liang, and V. Vinciotti, "An extended Kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, pp. 410–419, 7 2009.

[5] M. M. Mansouri, H. N. Nounou, M. N. Nounou, and A. A. Datta, "State and parameter estimation for nonlinear biological phenomena modeled by S-systems," *Digital Signal Processing*, vol. 28, pp. 1–17, 2014.

[6] G. Lillacci and M. Khammash, "Parameter estimation and model selection in computational biology," *PLoS Computational Biology*, vol. 6, 2010.

[7] N. J. Linden, B. Kramer, and P. Rangamani, "Bayesian parameter estimation for dynamical models in systems biology," *PLoS Computational Biology*, vol. 18, no. 10, p. e1010651, 2022.

[8] T. Kurdyaeva and A. Milias-Argeitis, "Uncertainty propagation for deterministic models of biochemical networks using moment equations and the extended Kalman filter," *Journal of the Royal Society Interface*, vol. 18, no. 181, p. 20210331, 2021.

[9] B. D. Anderson and J. B. Moore, *Optimal filtering*. New York: Dover Publications, 2005.

[10] R. Schneider and C. Georgakis, "How to not make the extended Kalman filter fail," *Industrial & Engineering Chemistry Research*, vol. 52, no. 9, pp. 3354–3362, 2013.

[11] A. Dey, K. Chakrabarti, K. K. Gola, and S. Sen, "A Kalman filter approach for biomolecular systems with noise covariance updating," in *2019 Sixth Indian Control Conference (ICC)*, pp. 262–267, IEEE, 2019.

[12] D. T. Gillespie, "The chemical Langevin equation," *The Journal of Chemical Physics*, vol. 113, no. 1, pp. 297–306, 2000.

[13] A. H. Jazwinski, *Stochastic processes and filtering theory*. New York: Dover Publications, 2007.

[14] A. Goldbeter and D. E. Koshland Jr, "An amplified sensitivity arising from covalent modification in biological systems.," *Proceedings of the National Academy of Sciences*, vol. 78, no. 11, pp. 6840–6844, 1981.

[15] A. Dey and S. Sen, "Describing function-based approximations of biomolecular systems," *IET Systems Biology*, vol. 12, pp. 93–100, 6 2018.

[16] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Transactions on Automatic Control*, vol. 15, no. 2, pp. 175–184, 1970.

[17] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. New Jersey: John Wiley & Sons, 2015.