Proceedings of the 31st Chinese Control Conference
July 25-27, 2012, Hefei, China

Determine Measurement Set for Parameter Estimation in Biological Systems Modelling

Hong Yue[1], Jianfang Jia[2]

1. Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK
E-mail: hong.yue@eee.strath.ac.uk

2. School of Information and Communication Engineering, North University of China, Taiyuan 030051, P. R. China
E-mail: jiajf2002@163.com

Abstract: Parameter estimation is challenging for biological systems modelling since the model is normally of high dimension, the measurement data are sparse and noisy, and the cost of experiments is high. Accurate recovery of parameters depend on the quantity and quality of measurement data. It is therefore important to know what measurements to be taken, when and how through optimal experimental design (OED). In this paper we present a method to determine the most informative measurement set for parameter estimation of dynamic systems, in particular biochemical reaction systems, such that the unknown parameters can be inferred with the best possible statistical quality using the data collected from the designed experiments. System analysis using matrix theory is introduced to examine the number of necessary measurement variables. The priority of each measurement variable is determined by optimal experimental design based on Fisher information matrix (FIM). The applicability and advantages of the proposed method are illustrated through an example of a signal pathway model.

Key Words: Measurement Set Selection, Optimal Experimental Design, Parameter Estimation, Biological Systems

1 Introduction

Most mechanistic mathematical models developed for biological and other systems contain adjustable or unknown parameters, the values of which can be estimated from observations. Parameter estimation is challenging for bioprocesses modelling [1] due to: (1) lack of quantitative measurements of dynamic response data and the measurement data is often corrupted with noise; (2) the complex nature of biological systems with high-dimensional, nonlinear and poorly understood dynamics. In general, performing experiments to obtain rich data is expensive and time-consuming for such systems. The problem of designing experiments to generate efficient measurement data is thus of particular importance. The term 'optimal experimental design (OED)' or 'design of experiment' refers to designing experiments in such a way that the parameters can be estimated from the resulting experimental data with the best possible statistical quality. This is a subject area of growing interests particularly in systems biology since huge experimental efforts are required in model development. Various methodologies have been developed and successfully applied to a broad range of systems [2–4]. Interested readers can find comprehensive reviews on experimental design and applications for general systems in [5, 6] and biological and biochemical systems in [7, 8].

The number of unknown parameters is often large for biological system models compared to the limited measurement data, which raises the issue of identifiability. The checking of identifiability is essential in employing parameter estimation techniques such as least squares estimation, maximum likelihood method and Bayesian estimation, etc [9]. Two types of identifiability are considered: the *a priori* structural identifiability and the *posteriori* practical identifiability. Structural (global) identifiability is concerned with the question of the theoretical uniqueness of solutions for a given model and experiment. A nonlinear system is said to be structurally (globally) identifiable if each set of parameter values yields unique output trajectories. This property guarantees that, under ideal conditions of noise-free observations and error-free model structure, the unknown parameters can be uniquely estimated from the designed input-output experiment [10]. The structural identifiability is a theoretical property of the model and a necessary condition for a successful parameter estimation, however, it is not sufficient to guarantee estimation accuracy in practice [11]. Additional problems commonly encountered in practice are sparse and noisy data, weak effect of unknown parameters on the measured output, etc., which should be addressed in practical identifiability analysis. The identifiability of a parameter estimation problem can be improved through well-designed experiments in general.

In order to produce and collect information-rich data, experimental design can be considered from two aspects. One is the design of input perturbations (type, level and duration of input signals), the other is to determine when and what kind of observations should be taken. Design parameters include level of initial conditions, which input and output variables to be taken, what sampling schedule to follow, etc. In this paper, OED is performed on choosing the most suitable set of observation variables for parameter estimation, also called *measurement set selection* in earlier publications [12, 13]. In measurement set selection, we need to consider not only the issue of identifiability in theory, but also the experimental restrictions in biology. For example, in a wet-lab environment, normally only a small number of protein concentrations can be simultaneously measured in a timely fashion. It is therefore important to determine which observables would provide more information for parameter estimation. Given a set of unknown parameters to be estimated, we attempt to investigate: (1) the best (minimum) number of measurement variables to be used; and (2) the set of measurement variables to be chosen.

This work is partly supported by National Natural Science Foundation of China (NSFC) under Grant 61004045, and Research Fund for the Doctoral Program of Higher Education of China (20091420120007).

The rest of the paper is organized as follows. In Section 2, the preliminaries on parameter estimation and model-based OED is briefly introduced. In Section 3, firstly the general dynamic model is reformulated to improve the computational efficiency and facilitate further analysis, then the method to determine the minimum measurement set is discussed using the matrix theory, and the priorities of state variables are calculated by model-based OED. Using a simplified IκB α-NF-κB signal pathway model as an example, the applicability of the design method to biological systems modelling is illustrated in Section 4. Finally the conclusions and discussions are given in Section 5.

2 Parameter Estimation and Experimental Design Preliminaries

Consider a general ordinary differential equation model to describe the dynamics of biological systems

$$\dot{\mathbf{X}}(t) = f\left(\mathbf{X}(t), \mathbf{p}, \boldsymbol{\omega}\right), \ \mathbf{X}(t_0) = \mathbf{X}_0 \tag{1}$$

$$\mathbf{Y}(t) = h\left(\mathbf{X}(t), \mathbf{p}\right) + \xi(t) \tag{2}$$

$\mathbf{X} \in \mathbb{R}^n$ is the state vector with initial condition $\mathbf{X}_0$ and n the number of the state variables. Each component of $\mathbf{X}$ is denoted as x_i, which normally stands for molecule concentrations in biochemical system models. $\mathbf{p} \in \mathbb{R}^m$ is the parameter vector with m the number of parameters. The components of $\mathbf{p}$ mostly refer to kinetic reaction rates. $f(\cdot)$ is a column nonlinear function for states transition, which is often derived from the underlying biochemical mechanisms. The vector $\boldsymbol{\omega}$ is introduced to represent the experimental design parameters. $\mathbf{Y} \in \mathbb{R}^r$ is the measurement output vector with $r(r \leq n)$ being the number of measurement variables, and $h(\cdot)$ the measurement function reflecting the choice of observables. The signal ξ is assumed to be independently and identically distributed, additive, zero-mean Gaussian noise. Parameter estimation for system (1)-(2) can be obtained by the least-square algorithm

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}\in\Theta} \sum_{l=1}^{N} \left(\mathbf{Y}(t_l) - \hat{\mathbf{Y}}(\hat{\mathbf{p}}, t_l)\right)^T \mathbf{Q}^{-1} \left(\mathbf{Y}(t_l) - \hat{\mathbf{Y}}(\hat{\mathbf{p}}, t_l)\right) \tag{3}$$

where $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are measurement output and model prediction output, respectively. $\mathbf{Q}$ is the measurement error covariance matrix, the subscript l indicates sampling time, N is the total number of sampling points in the dimension of time.

The Fisher information matrix (FIM) quantifies the information content of the measurement data for parameter estimation. For a nonlinear dynamic system, the FIM is a nonlinear function of the estimated parameters under the assumption that the measurement noise is independently and identically distributed with a zero-mean Gaussian distribution. Denote $\mathbf{X} = [x_1, x_2, \cdots, x_n]^T$, $\mathbf{p} = [p_1, p_2, \cdots, p_m]^T$, the local sensitivity matrix is described as

$$\mathbf{S} = \partial\mathbf{X}/\partial\mathbf{p} = (s_{ij}), \ \ s_{ij} = \partial x_i/\partial p_j \tag{4}$$

The FIM is represented as a function of local sensitivity matrix:

$$\mathrm{FIM}(\mathbf{p}, \boldsymbol{\omega}) = \sum_{l=1}^{N} \mathbf{S}^T(t_l, \mathbf{p}, \boldsymbol{\omega}) \mathbf{Q}^{-1} \mathbf{S}(t_l, \mathbf{p}, \boldsymbol{\omega}). \tag{5}$$

Under the assumption of additive zero-mean Gaussian noise in measurement, an OED problem can be written as a general optimization problem to read

$$\boldsymbol{\omega}^* = \arg\max_{\boldsymbol{\omega}\in\Omega} \Phi\left(\mathrm{FIM}(\mathbf{p}, \boldsymbol{\omega})\right). \tag{6}$$

Ω is the design space for the experimental design vector $\boldsymbol{\omega}$, $\Phi(\cdot)$ indicates the widely used alphabetical experimental design criteria that are normally scalar functions of FIM, such as A-optimal, maximizing $\mathrm{trace(FIM)}$; D-optimal, maximising $\det(\mathrm{FIM})$; E-optimal, minimizing $\lambda_{max}(\mathrm{FIM}^{-1})$, etc. Here $\mathrm{trace}(\cdot)$ and $\det(\cdot)$ are trace and determinant of a matrix, $\lambda_{max}(\cdot)$ is the maximum eigenvalue of a matrix. These criteria are related to the size and shape of the confidence hyper-ellipsoid for estimated parameters, and will give slightly different experimental design results when choosing different criteria. The design using any of the three criteria turns out to be a convex optimization problem when the FIM is an appropriate function of the experimental design parameters [14]. Problem (6) is in general an NP-hard problem, and the computational cost of the optimization problem depends on the complexity of the model structure/dynamics.

3 Measurement Set Selection

3.1 Dynamic Model with Unknown Parameters

For a system containing known and unknown parameters, the parameter vector $\mathbf{p}$ can be separated into two sets: $\boldsymbol{\eta} \in \mathbb{R}^l$ for known parameters, and $\boldsymbol{\theta} \in \mathbb{R}^q$ for unknown parameters with $l+q=m$. Here it is reasonable to assume that the model is linear in parameters, as widely applied to biochemical systems taking kinetic rate coefficients as parameters to describe the individual reactions in a model. Considering a simple example of a generic reaction $S_1 + S_2 \xrightarrow{k} P$, the reaction rate is given by $k[S_1]^a[S_2]^b$ with $[\cdot]$ being the concentration of reaction species, and a, b reaction orders with respect to S_1 and S_2, respectively. k is the rate constant that is a linear term in describing the reaction rate. Under this assumption and together with the separation of the known and unknown terms in $\mathbf{p}$, model (1) can be further written as follows (for simplicity, $\boldsymbol{\omega}$ is omitted):

$$\dot{\mathbf{X}}(t) = g\left(\mathbf{X}(t)\right)\boldsymbol{\eta} + \varphi\left(\mathbf{X}(t)\right)\boldsymbol{\theta} \tag{7}$$

where $g(\cdot) \in \mathbb{R}^{n\times l}$ and $\varphi(\cdot) \in \mathbb{R}^{n\times q}$ are nonlinear functions associated with known and unknown parameters, respectively. For a biochemical system, the nonlinear function $g(\cdot)$ often contains both linear and nonlinear terms with respect to species concentrations (state variables). A typical nonlinear form involving two reaction species is a bilinear function. When a system has a large number of reactions, leading to a high dimension in model parameters, the separation of the linear (states) terms from the nonlinear (states) terms will decompose the model into subgroups with a reduced size in each group. This will largely improve the efficiency of numerical calculations that often involve integration operation of matrix functions. Following this idea, model (7) is further reformulated to be:

$$\dot{\mathbf{X}}(t) = \mathbf{A}\mathbf{X}(t) + \tilde{g}\left(\mathbf{X}(t)\right)\boldsymbol{\eta}_1 + \varphi\left(\mathbf{X}(t)\right)\boldsymbol{\theta} \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a parameter matrix, $\tilde{g}(\cdot)$ groups the nonlinear (states) functions in $g(\cdot)$, $\boldsymbol{\eta}_1 \in \mathbb{R}^{l_1}(l_1 \leq l)$ is

the known parameter vector associated with $\tilde{g}(\cdot)$. Note that with this new formulation that isolate the unknown parameters from the whole parameter set, the term $\mathbf{p}$ of the FIM function in (6) should be replaced by $\boldsymbol{\theta}$ in OED.

3.2 Minimum State Number to be Measured

A general assumption is made that *measurement output* $\mathbf{Y}$ are linear function of the states. This is how measurement data is processed with most current measurement techniques applied to biological or biochemical systems. The measurement output in (2) can then be written as (ignoring the noise term for simplicity)

$$\mathbf{Y}(t) = \mathbf{C}\mathbf{X}(t) \tag{9}$$

where $\mathbf{C} \in \mathbb{R}^{r\times n}$ is the measurement matrix. From model (8) and (9), the output reads

$$\begin{aligned}\mathbf{Y}(t) &= \mathbf{C}e^{\mathbf{A}t}\mathbf{X}_0 + \mathbf{C}\left(\int_0^t e^{\mathbf{A}(t-\tau)}\tilde{g}(\mathbf{X}(\tau))d\tau\right)\boldsymbol{\eta}_1 \\ &\quad + \mathbf{C}\left(\int_0^t e^{\mathbf{A}(t-\tau)}\varphi(\mathbf{X}(\tau))d\tau\right)\boldsymbol{\theta}\end{aligned} \tag{10}$$

Equation (10) shows the linear dependency of measurement observables on unknown parameters $\boldsymbol{\theta}$. According to the linear matrix theory, the rank of the linear term multiplied to $\boldsymbol{\theta}$, i.e. $\operatorname{rank}\left(\mathbf{C}\int_0^t e^{\mathbf{A}(t-\tau)}\varphi(\mathbf{X}(\tau))d\tau\right)$ should be maximised in order to realise the minimum number of measurement variables for the estimation of $\boldsymbol{\theta}$. The design problem can then be formulated as an optimisation problem of choosing a matrix $\mathbf{C}$, consisting of elements 1 or 0, so as to maximise the following objective function:

$$J(\mathbf{C}) = \max_{\mathbf{C}} \operatorname{rank}\left(\mathbf{C}\int_0^t e^{\mathbf{A}(t-\tau)}\varphi(\mathbf{X}(\tau))d\tau\right) \tag{11}$$

The solution to (11) is discussed in the following. Denote

$$\mathbf{B} = \int_0^t e^{\mathbf{A}(t-\tau)}\varphi(\mathbf{X}(\tau))d\tau \tag{12}$$

where $\mathbf{B} \in \mathbb{R}^{n\times q}$ represents the convolution of $e^{\mathbf{A}(t-\tau)}$ and $\varphi(\mathbf{X}(\tau))$. For a given model, the matrix term $\mathbf{A}$ and function $\varphi(\cdot)$ are known, therefore $\mathbf{B}$ can be taken as a known term at time t. Assume that $\operatorname{rank}(\mathbf{B}) = m$, from matrix theory it is known that $\operatorname{rank}(\mathbf{CB}) \leq min\{\operatorname{rank}(\mathbf{C}), \operatorname{rank}(\mathbf{B})\}$, which means $J(\mathbf{C})$ won't be larger than m in any case. The conclusion is therefore made that $\max J(\mathbf{C}) = m$ when $\operatorname{rank}(\mathbf{C}) = m$.

It should be noted that the minimum number of observables determined this way is a theoretical result that guarantees the structural identifiability and the best estimation accuracy. Parameter estimation in practice is not restricted to the minimum number of measurement variables but the estimation result is only an approximate solution.

3.3 Priority of Measurement Variables

As denoted in the general nonlinear model of the dynamic systems (1)-(1), there are n state variables and each of them can be taken as the observables via the measurement matrix $\mathbf{C}$. To prioritise each variable x_i in terms of their contributions to the specified parameter estimation problem, the weighting factor ω_i is introduced to x_i to form the design problem.

$$\zeta = \begin{pmatrix} x_1, x_2, \cdots, x_n \\ \omega_1, \omega_2, \cdots, \omega_n \end{pmatrix}, \ \sum_{i=1}^{n}\omega_i = 1, \ \omega_i \geq 0 \tag{13}$$

Taking the design parameter vector as $\boldsymbol{\omega} = [\omega_1, \omega_2, \cdots, \omega_n]^T$, computationally the FIM can be written as

$$\mathrm{FIM}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{l}^{N}\sum_{i=1}^{n}\omega_i \mathbf{S}_i^T(t_l, \boldsymbol{\theta})\mathbf{S}_i(t_l, \boldsymbol{\theta}) \tag{14}$$

where $\mathbf{S}_i$ is the ith row of the sensitivity matrix $\mathbf{S}$.

The idea of the E-optimal design is to minimise the largest confidence interval of the estimated parameters. Taking this criterion, the OED problem on measurement set selection is formulated as follows:

$$\begin{aligned}\boldsymbol{\omega}^* &= \arg\min_{\boldsymbol{\omega}\in\Omega} \lambda_{max}\left[(\mathrm{FIM}(\boldsymbol{\theta}, \boldsymbol{\omega}))^{-1}\right] \\ s.t. &\sum_{i=1}^{n}\omega_i = 1, \ \omega_i \geq 0\end{aligned} \tag{15}$$

This problem can be recast into a semidefinite program (SDP) [13, 15]:

$$\begin{aligned}\boldsymbol{\omega}^* &= \arg\max_{\boldsymbol{\omega}} \nu \\ s.t. \ &\sum_{i=1}^{n}\omega_i \mathbf{S}_i^T(t_l, \boldsymbol{\theta})\mathbf{S}_i(t_l, \boldsymbol{\theta}) \geq \nu \mathbf{I}_q \\ &\sum_{i=1}^{n}\omega_i = 1, \ \omega_i \geq 0\end{aligned} \tag{16}$$

$\mathbf{I}_q$ is the $q \times q$ identity matrix. The optimisation can then be solved efficiently by many SDP solvers such as SeDuMi, a high quality package with MATLAB interface.

4 Simulation Study on IκB-NF-κB Signalling Pathway Model

4.1 Model Simulation and E-optimal Design Result

To examine the applicability of this method in parameter estimation of biological models, a simplified IκB-NF-κB signal transduction pathway network model is chosen for simulation study. The protein NF-κB is a fundamental component of the IκB-NF-κB signaling pathway that regulates numerous genes [16], acting in response to environmental and biological stress, and bacterial and viral infection. Its specificity and its role in the temporal control of gene expressions are of crucial physiological interest. The mechanism of this pathway has been described by Hoffmann et al. [17], Nelson et al.[18], Lipniacki et al.[19] and Ashall et al.[20], to name a few.

The simplified model is written in a set of ordinary differential equations with 10 state variables and 24 parameters (see appendix for more details). This model is linear in parameters, but the dynamic transition function contains linear terms, bilinear terms and a quadratic term. From our previous work of global sensitivity analysis of this model [21], a set of five parameters are identified to be the most sensitive ones and they are thus used as the unknown parameters in the

simulation study. The five unknown parameters are written in a vector format as $\boldsymbol{\theta} = \begin{bmatrix} \theta_5 & \theta_{12} & \theta_{13} & \theta_{16} & \theta_{18} \end{bmatrix}^T$. To improve the calculation efficiency, we first rewrote the model into the format of (8) and have obtained $q = 5, l = 19, l_1 = 6$, $\boldsymbol{\eta}_1 = \begin{bmatrix} \theta_1 & \theta_3 & \theta_9 & \theta_{11} & \theta_{14} & \theta_{20} \end{bmatrix}^T$. The objective of OED is to select the most informative state variables from the 10 states to provide the best estimation accuracy for the 5 unknown parameters.

In the simulation, the nominal values of the five parameters are $\boldsymbol{\theta}^* = [1.221\ 0.99\ 0.0168\ 0.2448\ 0.018]$, the initial conditions of the states were taken from the equilibrium with $x_6 = 0.1\mu M$ as an activation input (IKK). A Gaussian noise was introduced into the simulation data with zero-mean and a standard deviation of 1 % of the 'clean' signal at each time point. For large-scale biological models, due to limitations in experimental measurement frequency, the measured data are often sampled at relatively large time spans. In this numerical study, the sampling points are taken between 0 and 360 minutes with 5 minutes being the sampling interval. It is also assumed that each protein concentration (state variable) can be measured independently in the experiment. The E-optimal design was calculated over an uncertainty region around the nominal values [13], and the state variables in descending order of priority are presented as follows:

$$\mathbf{X}^* = [x_5\ x_8\ x_7\ x_1\ x_{10}\ x_4\ x_3\ x_9\ x_2\ x_6].$$

This OED result indicates that, for the 5 unknown parameters to be estimated, among the 10 state variables, x_5 is the most informative measurement variable, x_8 is the second informative one and so on and so forth. When selecting the measurement set for parameter estimation, we should consider those states with higher priorities so as to obtain a higher estimation accuracy.

4.2 Discussions on Measurement Set Selection

From the IκB-NF-κB signalling pathway differential equation model, we wrote the parameter matrix $\mathbf{A}$ and function $\varphi(\cdot)$ following (8). Accordingly, the rank of the matrix $\mathbf{B}$ in (12) was computed by the convolution integration and this calculation brings $\mathrm{rank}(\mathbf{B}) = 5$. Following the discussions in Section 3.2, when $\mathrm{rank}(\mathbf{C}) = \mathrm{rank}(\mathbf{B}) = 5$, $\max J(\mathbf{C}) = 5$, which means the minimum number of the measurement states is 5 to guarantee the structure identifiability in estimating $\boldsymbol{\theta}$. This result is intuitive since there are 5 (independent) unknown parameters to be estimated and all the state variables are measured independently. Taking into account the E-optimal experiment design result in $\mathbf{X}^*$, we can select the top five states $[x_5\ x_8\ x_7\ x_1\ x_{10}]$ to form the most suitable measurement set.

To investigate how the measurement set selection may affect the parameter estimation, the following four experiments taking different state variables are implemented for comparison.

(a) 3 top observables in $\mathbf{X}^*$, $[x_5\ x_8\ x_7]$;
(b) 5 top observables in $\mathbf{X}^*$, $[x_5\ x_8\ x_7\ x_1\ x_{10}]$;
(c) 7 top observables in $\mathbf{X}^*$, $[x_5\ x_8\ x_7\ x_1\ x_{10}\ x_4\ x_3]$;
(d) 5 bottom observables in $\mathbf{X}^*$, $[x_4\ x_3\ x_9\ x_2\ x_6]$.

In the first 3 experiments, the number of observables is different in each case but the measurement states are always selected from the top following the ranking given in $\mathbf{X}^*$. In the last experiment, the number of observables is taken as the minimum number but a different set of measurement variables were selected. The least-square algorithm was used for parameter estimation, in which the parameter searching space in all simulations were set to be $[0.01\boldsymbol{\theta}^*, 10\boldsymbol{\theta}^*]$, and the initial searching point was randomly chosen within the parameter space. Multi-shooting strategy was employed to avoid the local minimum problem. The estimated parameter values are given in Table 1. All estimations bring reasonable recovery of the parameter values, among them the results using 5 and 7 optimal measurement variables have less estimation errors than those using 3 optimal observables or 5 non-optimal observables.

Table 1: Estimated Parameters with Different Observables

	$\hat{\theta}_5$	$\hat{\theta}_{12}$	$\hat{\theta}_{13}$	$\hat{\theta}_{16}$	$\hat{\theta}_{18}$
(a)	1.181	0.955	0.0162	0.2361	0.0174
(b)	1.209	0.978	0.0166	0.2419	0.0178
(c)	1.209	0.978	0.0166	0.2428	0.0178
(d)	1.158	0.936	0.0159	0.2316	0.0170

Since the result of parameter estimation highly relies on the efficiency of the optimisation algorithm, it is perhaps not the best way to evaluate the effects of measurement set selection. Confidence interval, instead, is a more reliable assessment regarding each design and is worked out from the FIM following Cramer-Rao inequality. In general, a smaller confidence interval indicates an estimation with less errors, and vice versa. For the first 3 experiments, the corresponding 95% confidence interval of several parameter pairs are illustrated in Fig. 1 to Fig. 4, in which $'+'$ stands for the nominal value of the parameters. Two parameters are chosen in each figure just to present the results in a 2D plane.

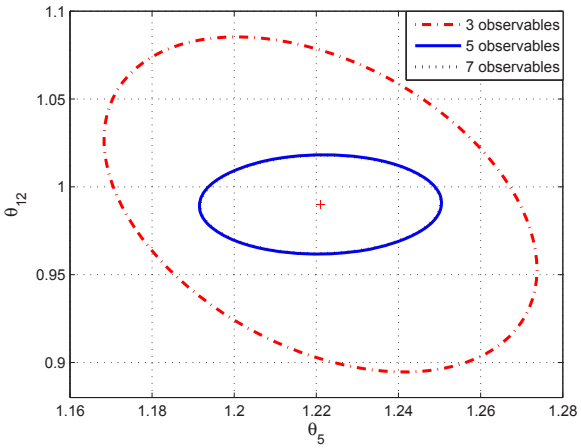

Fig. 1: Confidence interval of parameters θ_5 and θ_{12}

It can be seen from Fig. 1 to Fig. 4 that, for the case of three optimal observables, the 95% confidence interval is much larger than that of the five or seven optimal observables. Whereas, for the experiments with five or more observables, their 95% confidence intervals are very close to each other, in fact, the ellipsoids are visually indistinguishable in Fig. 1 to Fig. 4. This result suggests that when the number of measurement variables used is less than the min-

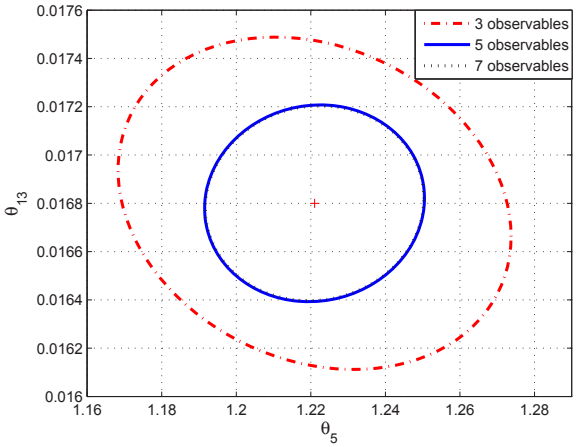

Fig. 2: Confidence interval of parameters θ_5 and θ_{13}

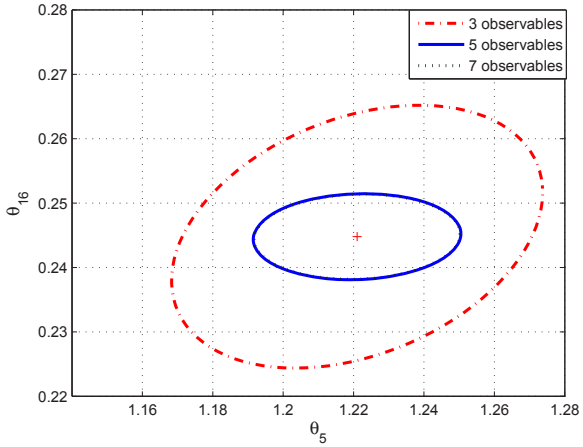

Fig. 3: Confidence interval of parameters θ_5 and θ_{16}

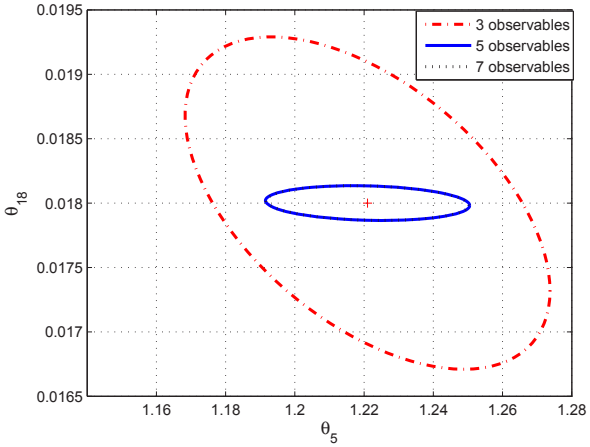

Fig. 4: Confidence interval of parameters θ_5 and θ_{18}

imum number of states to be measured, the estimation accuracy could be poor even when the most informative state variables are selected. Certain information about the unknown parameters set $\boldsymbol{\theta}$ are missing when using less than necessary measurements. On the other hand, the estimation results won't improve much when more than necessary measurements are taken into calculation. This is also validated by the parameter estimation results in Table 1.

When selecting measurement set, it is also important to take the more informative observables rather than those containing less information. By comparing the confidence interval ellipsoids in Fig. 5, it can be clearly seen that the confidence interval using the 5 optimal observables (top 5 states in $\mathbf{X}^*$) is much smaller than the one using 5 non-optimal observables (bottom 5 states in $\mathbf{X}^*$). The former has a smaller parameter estimation error owing to the fact that the selected measurement set contains more information about the unknown parameters.

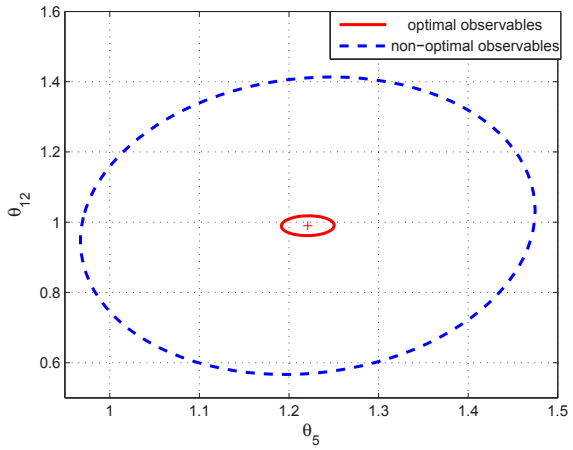

Fig. 5: Comparison of confidence interval of parameters θ_5 and θ_{12} w.r.t. the optimal and non-optimal observables

5 Conclusions and Discussions

Optimally designed experiments allow to maximise the information contained in measurement data and also to minimise cost and efforts of experiments. In this work, the measurement set selection problem is discussed where the number of measurement variables and the priority of observables can be determined through matrix theory and model-based OED. In the case study example, it is assumed that each state variable can be measured independently. Therefore, the result on the minimum number of state variables to be measured is quite intuitive. In some practical problems, only the combination of states can be measured rather than each individual state. In such cases, the proposed method still applies since the priority of any combined state measurements can be extracted from the ranking or weights of each individual state variable. Also, the minimum number of states to be measured can still be calculated by the proposed method using matrix theory. We are interested in exploring such examples from biological or biochemical systems, and further validate and develop the measurement set selection strategy.

The OED on measurement set selection will be particularly useful when the number of observables available is large or when new observables such as new antibodies are to be generated. To some extend, experimental design bridges the gap between theoretical and experimental research communities. On the one hand, theoreticians learn to evalu-

ate and appreciate feasibility and efforts required in experiments, on the other hand, experimental scientists develop a better understanding on which kind of information is most helpful to model development.

Acknowledgement

We would like to thank Dr. Taiyuan Liu for the helpful discussions and aid in programming, also thank Dr. Fei He for helping with the programming of experimental design and confidence interval output.

Appendix

The model presented here is a simplified version of the NF-κB signal pathway model [17] with IκBβ and IκBε knock out. The reaction species and state variable definition is given in Table 2, in which the subscript '-t' represents the mRNA corresponding to the former protein and 'n' indicates the proteins inside nucleus. The values of model parameters are listed in Table 3 with units of μM for concentration and minute for time. The constant term Source is taken to be 1μM in ODEs.

Table 2: IκB-NF-κB Model States

States	Species	States	Species
x_1	IκBα	x_6	IKK
x_2	NF-κB	x_7	NF-κB$_n$
x_3	IκBα-NF-κB	x_8	IκBα_n
x_4	IKKIκBα	x_9	IκBα_n-NF-κB$_n$
x_5	IKKIκBα-NF-κB	x_{10}	IκBα_{-t}

Table 3: IκB-NF-κB Model Parameter Values

θ_1	30	θ_9	30	θ_{17}	0.00678
θ_2	6e-5	θ_{10}	6e-5	θ_{18}	0.018
θ_3	30	θ_{11}	9.24e-5	θ_{19}	0.012
θ_4	6e-5	θ_{12}	0.99	θ_{20}	11.1
θ_5	1.221	θ_{13}	0.0168	θ_{21}	0.075
θ_6	6e-5	θ_{14}	1.35	θ_{22}	0.828
θ_7	5.4	θ_{15}	0.075	θ_{23}	0.0072
θ_8	0.0048	θ_{16}	0.2448	θ_{24}	0.2442

A set of ordinary differential equations are used to describe the system dynamics.

$$
\begin{aligned}
\dot{x}_1 &= (\theta_{17}+\theta_{18})x_1 + \theta_2 x_3 + \theta_{15}x_4 + \theta_{19}x_8 + \theta_{16}x_{10} \\
&\quad -\theta_1 x_1 x_2 - \theta_{14} x_1 x_6 \\
\dot{x}_2 &= -\theta_7 x_2 + (\theta_2+\theta_6)x_3 + (\theta_4+\theta_5)x_5 + \theta_8 x_7 \\
&\quad -\theta_1 x_1 x_2 - \theta_3 x_2 x_4 \\
\dot{x}_3 &= -(\theta_2+\theta_6)x_3 + \theta_{21}x_5 + \theta_{22}x_9 + \theta_1 x_1 x_2 \\
&\quad -\theta_{20} x_3 x_6 \\
\dot{x}_4 &= -(\theta_{15}+\theta_{24})x_4 + \theta_4 x_5 + \theta_{14} x_1 x_6 - \theta_3 x_2 x_4 \\
\dot{x}_5 &= -(\theta_4+\theta_5+\theta_{21})x_5 + \theta_3 x_2 x_4 + \theta_{20} x_3 x_6 \\
\dot{x}_6 &= (\theta_{15}+\theta_{24})x_4 + (\theta_5+\theta_{21})x_5 - \theta_{23}x_6 - \theta_{14}x_1 x_6 \\
&\quad -\theta_{20} x_3 x_6 \\
\dot{x}_7 &= \theta_7 x_2 - \theta_8 x_7 + \theta_{10} x_9 - \theta_9 x_7 x_8 \\
\dot{x}_8 &= \theta_{18} x_1 - \theta_{19} x_8 + \theta_{10} x_9 - \theta_9 x_7 x_8 \\
\dot{x}_9 &= -(\theta_{10}+\theta_{22})x_9 + \theta_9 x_7 x_8 \\
\dot{x}_{10} &= \theta_{11} Source - \theta_{13} x_{10} + \theta_{12} x_7^2
\end{aligned}
$$

References

[1] E.O. Voit, *Computational Analysis of Biochemical Systems*. Cambridge, UK: Cambridge University Press, 2000.

[2] A.C. Atkinson, A.N. Donev, and R. Tobias, *Optimum Experimental Designs, with SAS*, Oxford University Press, 2007.

[3] D.C. Montomery, *Design and Analysis of Experiments*, 5th ed. New York: John Wiley, 2001.

[4] P.E. Box, J.S. Hunter, and W.G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed. New Jersey: Wiley Insterscience, 2005.

[5] K. Chaloner and I. Verdinelli, Bayesian experimental design: a review, *Statist. Sci.*, 10(3): 273-304, 1995.

[6] L. Pronzato, Optimal experimental design and some related control problems, *Automatica*, 44(2): 303-325, 2008.

[7] G. Franceschini and S. Macchietto, Model-based design of experiments for parameter precision: State of the art, *Chemical Engineering Science*, 63(19): 4846-4872, 2008.

[8] C. Kreutz and J. Timmer, Systems biology: experimental design, *FEBS Journal*, 276(4): 923-942, 2009.

[9] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice Hall, 1999.

[10] S. Audoly, G. Bellu, L. DAngio, M.P. Saccomani and C. Cobelli, Global identifiability of nonlinear models of biological systems, *IEEE Trans. on Biomedical Engineering*, 48(1): 55-65, 2001.

[11] M. Rodriguez-Fernandez, P. Mendes, and J.R. Banga, A hybrid approach for efficient and robust parameter estimation in biochemical pathways, *BioSystems*, 83(2-3): 248-265, 2006.

[12] H. Yue, M. Brown, F. He, J.F. Jia and D.B. Kell, Sensitivity analysis and robust experimental design of a signal transduction pathway system, *International Journal of Chemical Kinetics*, 40(11): 730-741, 2008.

[13] F. He, M. Brown, and H. Yue, Maximin and Bayesian robust experimental design for measurement set selection in modelling biochemical regulatory systems, *International Journal of Robust and Nonlinear Control*, 24(6): 1059-1078, 2010.

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

[15] P. Flaherty, M.I. Jordan, and A.P. Arkin, Robust design of biological experiments, in *Proc. of the Neural Information Processing Systems (NIPS)*, Cambridge, USA, 2006: 363-370.

[16] N.D. Perkins, Integrating cell-signalling pathways with NF-κB and IKK function, *Nature Reviews Molecular Cell Biology*, 8(1): 49-62, 2007.

[17] A. Hoffmann, A. Levchenko, M.L. Scott and D. Baltimore, The IκB-NF-κB signaling module: temporal control and selective gene activation, *Science*, 298, 1241-1245, 2002.

[18] D.E. Nelson, A.E.C. Ihekwaba, M. Elliott, J. Johnson, C.A. Gibney, B.E. Foreman, G. Nelson, V. See, C.A. Horton, D.G. Spiller, S.W. Edwards, H.P. McDowell, J.F. Unitt, E. Sullivan, R. Grimley, N. Benson, D. Broomhead, D.B. Kell, and M.R.H. White, Oscillations in NF-κB signaling control the dynamics of gene expression, *Science*, 306: 704-708, 2004.

[19] T. Lipniacki, P. Paszek, A.R. Brasier, B. Luxon, and M. Kimmel, Mathematical model of NF-κB regulatory module, *Journal of Theoretical Biology*, 228(2):195-215, 2004.

[20] L. Ashall, C.A. Horton, D.E. Nelson, P. Paszek, C.V. Harper, K. Sillitoe, S. Ryan, D.G. Spiller, J.F. Unitt, D.S. Broomhead, D.B. Kell, D.A. Rand, V. Sée, and M.R.H. White, Pulsatile stimulation determines timing and specificity of NF-κB-dependent transcription, *Science*, 324(5924), 242-246, 2009.

[21] Y.S. Jin, H. Yue, M. Brown, Y. Liang, and D.B. Kell, Improving data fitting of a signal transduction model by global sensitivity analysis, in *Proc. American Control Conference*, New York, USA, 2007: 2708-2713.