Article ID:0253-2778(2012)10--

Measurement set selection of parameter estimation in biological systems modelling — A case study for signal transduction pathway

JIA Jianfang[1], YUE Hong[2]

(1. *School of Information and Communication Engineering, North University of China, Taiyuan* 030051, *China*;
2. *Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow* G1 1XW, UK)

Abstract: Parameter estimation is a challenging problem 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 depends on the quantity and quality of measurement data. It is therefore important to know which measurements to be taken, when and how through optimal experimental design (OED). In this paper a method was proposed 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 was used to examine the number of necessary measurement variables. The priority of each measurement variable was determined by optimal experimental design based on Fisher information matrix (FIM). The applicability and advantages of the proposed method were shown through an example of a signal pathway model.

Key words: measurement set selection; optimal experimental design; parameter estimation; biological systems

CLC number: N945.14; N941 **Document code**: A doi:10.3969/j.issn.0253-2778.2012.10.00

Citation: Jia Jianfang, Yue Hong. Measurement set selection of parameter estimation in biological systems modelling — A case study for signal transduction pathway[J]. Journal of University of Science and Technology of China, 2012, 42(10): 100-105.

生物系统建模中参数估计的测量集选择——以信号转导通路模型研究为例

贾建芳[1], 岳 红[2]

(1. 中北大学信息与通信工程学院, 山西太原 030051; 2. 斯特莱斯克莱德大学电子与电气工程系, 英国格拉斯哥 G1 1XW)

摘要: 生物系统模型通常具有很高的维数, 测量数据不完备、易受噪声污染, 而且生物实验成本高, 所以参数

Received: 2012-04-16; **Revised**: 2012-05-17

Foundation item: Supported by National Natural Science Foundation of China (NSFC) (61004045), and Research Fund for the Doctoral Program of Higher Education of China (20091420120007).

Biography: JIA Jianfang, male, born in 1973, PhD/associate Prof. Research field: biological system. E-mail: jiajf2002@163.com

Corresponding author: YUE Hong, PhD. E-mail: hong.yue@eee.strath.ac.uk

The earlier version has been accepted for 2012 *Chinese Control Conference.*

估计已经成为生物系统建模的挑战性问题之一.参数的精确估计取决于测量数据的数量和质量,因此,通过优化实验设计确定如何采集测量数据是非常重要的.针对动态系统的参数估计问题,尤其是生物反应系统,提出了一种确定富含信息的测量集选择方法,通过从设计的实验中获得测量数据,以最佳的统计质量估计系统的未知参数.该方法首先利用矩阵论的系统分析来确定估计参数所必需的测量状态的数目,再通过基于Fisher信息阵的优化实验设计决定每个测量状态的优先等级.最后,以信号转导通路模型为例,解释了该方法的优势和适用性.

关键词:测量集选择;优化实验设计;参数估计;生物系统

0 Introduction

Model-based analysis of complex biological networks is a major topic of current systems biology. Most mechanistic mathematical models developed for biological and other systems contain adjustable or unknown parameters, the values of which can be estimated from observations. The dynamics of the model often is sensitive to parameters especially for the oscillation system, whether the model parameters are estimated optimally or not will directly affect the performance of the model. As a result, parameter estimation is challenging for bioprocesses modelling[1]: 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 are expensive and time-consuming for such systems. The problem of designing experiments to generate efficient measurement data is thus of particular importance. Optimal experimental design (OED) 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 Refs. [5-6] and biological and biochemical systems in Refs. [7-8].

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. The identifiability of a parameter estimation problem can be improved through well-designed experiments in general. In this paper, OED was performed on choosing the most suitable set of observation variables for parameter estimation, also called measurement set selection in earlier publications[9-10]. 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 observable 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.

The rest of the paper is organized as follows. In Section 1, the preliminaries on parameter estimation and model-based OED is briefly introduced. In Section 2, 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 3. Finally the conclusions and discussions are given in Section 4.

1 Parameter estimation and experimental design preliminaries

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

$$\dot{\boldsymbol{X}}(t) = f(\boldsymbol{X}(t), \boldsymbol{p}, \boldsymbol{\omega}), \boldsymbol{X}(t_0) = \boldsymbol{X}_0 \quad (1)$$

$$\boldsymbol{Y}(t) = h(\boldsymbol{X}(t), \boldsymbol{p}) + \xi(t) \quad (2)$$

$\boldsymbol{X} \in \mathbb{R}^n$ is the state vector with initial condition $\boldsymbol{X}_0$ and n the number of state variables. Each component of $\boldsymbol{X}$ is denoted as x_i, which normally stands for molecule concentrations in biochemical system models. $\boldsymbol{p} \in \mathbb{R}^m$ is the parameter vector with m the number of parameters. The components of $\boldsymbol{p}$ 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. $\boldsymbol{Y} \in \mathbb{R}^r$ is the measurement output vector with $r(r \leqslant n)$ 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

$$\hat{\boldsymbol{p}} = \arg \min_{\boldsymbol{p} \in \Theta} \sum_{l=1}^{N} (\boldsymbol{Y}(t_l) - \hat{\boldsymbol{Y}}(\hat{\boldsymbol{p}}, t_l))^{\mathrm{T}} \boldsymbol{Q}^{-1} \cdot (\boldsymbol{Y}(t_l) - \hat{\boldsymbol{Y}}(\hat{\boldsymbol{p}}, t_l)) \quad (3)$$

where $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ are measurement output and model prediction output, respectively. $\boldsymbol{Q}$ 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.

Denote $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]^{\mathrm{T}}$, $\boldsymbol{p} = [p_1, p_2, \cdots, p_m]^{\mathrm{T}}$, the local sensitivity matrix is described as

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

The Fisher information matrix (FIM) is represented as a function of local sensitivity matrix:

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

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

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

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 trace (FIM); D-optimal, maximising det (FIM); E-optimal, minimizing $\lambda_{\max}(\mathrm{FIM}^{-1})$, etc. Here 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[11]. 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.

2 Measurement set selection

2.1 Dynamic model with unknown parameters

For a system containing known and unknown parameters, the parameter vector $\boldsymbol{p}$ 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. Under this assumption,

together with the separation of the known and unknown terms in $\boldsymbol{p}$, model (1) can be further written as follows (for simplicity, $\boldsymbol{\omega}$ is omitted)

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

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. For a biochemical system, the nonlinear function $g(\cdot)$ often contains both linear and nonlinear terms with respect to species concentrations (state variables). 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

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

where $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is a parameter matrix, $\tilde{g}(\cdot)$ groups the nonlinear (states) functions in $g(\cdot)$, $\boldsymbol{\eta}_1\in\mathbb{R}^{l_1}$ ($l_1\leqslant l$) is the known parameter vector associated with $\tilde{g}(\cdot)$. Note that using this new formulation to isolate the unknown parameters from the whole parameter set, the term $\boldsymbol{p}$ of the FIM function in Eq. (6) should be replaced by $\boldsymbol{\theta}$ in OED.

2.2 Minimum state number to be measured

A general assumption is made that measurement output $\boldsymbol{Y}$ 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 system (2) can then be written as (ignoring the noise term for simplicity)

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

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

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

Eq. (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. $\mathrm{rank}\left(\boldsymbol{C}\int_0^t \mathrm{e}^{\boldsymbol{A}(t-\tau)}\varphi(\boldsymbol{X}(\tau))\mathrm{d}\tau\right)$ 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 $\boldsymbol{C}$, consisting of elements 1 or 0, so as to maximise the following objective function

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

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

Denote

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

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

It should be noted that the minimum number of observables determined by 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 and the estimation result is only an approximate solution.

2.3 Priority of measurement variables

As denoted in the general nonlinear model of the dynamic systems (1)～(2), there are n state variables and each of them can be taken as the observables via the measurement matrix $\boldsymbol{C}$. 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.

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

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

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

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

$$\boldsymbol{\omega}^* = \arg \min_{\boldsymbol{\omega} \in \Omega} \lambda_{\max}[(\mathrm{FIM}(\boldsymbol{\theta}, \boldsymbol{\omega}))^{-1}] \quad (15)$$

$$\text{s. t.} \quad \sum_{i=1}^{n} \omega_i = 1, \omega_i \geqslant 0$$

This problem can be recast into a semi-definite program (SDP)[10,12]:

$$\boldsymbol{\omega}^* = \arg \max_{\boldsymbol{\omega}} \nu \quad (16)$$

$$\text{s. t.} \quad \sum_{i=1}^{n} \omega_i \mathbf{S}_i^{\mathrm{T}}(t_l, \boldsymbol{\theta}) \mathbf{S}_i(t_l, \boldsymbol{\theta}) \geqslant \nu \mathbf{I}_q$$

$$\sum_{i=1}^{n} \omega_i = 1, \omega_i \geqslant 0$$

$\mathbf{I}_q$ 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.

3 Simulation study on IκB-NF-κB signalling pathway model

3.1 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 was chosen for simulation study. The reaction species and state variable definition is given in Tab. 1, 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 Tab. 2 with units of μM for concentration and minute for time. The constant term Source is taken to be 1 μM in ODEs.

Tab. 1　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$_{-t}$

Tab. 2　IκB-NF-κB model parameter values

parameter	value	parameter	value	parameter	value
θ_1	30	θ_9	30	θ_{17}	0.006 78
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}	0.016 8	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}	0.007 2
θ_8	0.0048	θ_{16}	0.244 8	θ_{24}	0.244 2

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

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

From our previous work of global sensitivity analysis of this model[18], a set of five parameters are identified to be the most sensitive ones and they are thus used as the unknown parameters vector $\boldsymbol{\theta}=[\theta_5 \quad \theta_{12} \quad \theta_{13} \quad \theta_{16} \quad \theta_{18}]^{\mathrm{T}}$ in the simulation study. To improve the calculation efficiency, we first rewrote the model into the format of (8) and have obtained $q=5$, $l=19$, $l_1=6$, $\boldsymbol{\eta}_1=[\theta_1 \quad \theta_3 \quad \theta_9 \quad \theta_{11} \quad \theta_{14} \quad \theta_{20}]^{\mathrm{T}}$. 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 $\boldsymbol{\theta}^* = [1.221 \quad 0.99 \quad 0.0168 \quad 0.244\,8 \quad 0.018]$, 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. 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[10], and the state variables in descending order of priority are presented as follows

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

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. When selecting the measurement set for parameter estimation, we should consider those states with higher priorities so as to obtain a higher estimation accuracy.

3.2 Discussions on measurement set selection

From the IκB-NF-κB signalling pathway differential equation model, we wrote the parameter matrix $\boldsymbol{A}$ and function $\varphi(\cdot)$ following (8). Accordingly, the rank of the matrix $\boldsymbol{B}$ in (12) was computed by the convolution integration and this calculation brings rank($\boldsymbol{B}$)=5. Following the discussions in Section 2.2, when rank($\boldsymbol{C}$) = rank($\boldsymbol{B}$)=5, max$J$($\boldsymbol{C}$) = 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 $\boldsymbol{X}^*$, we can select the top five states $[x_5 \quad x_8 \quad x_7 \quad x_1 \quad x_{10}]$ 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.

① 3 top observables in $\boldsymbol{X}^*$, $[x_5 \quad x_8 \quad x_7]$;

② 5 top observables in $\boldsymbol{X}^*$, $[x_5 \quad x_8 \quad x_7 \quad x_1 \quad x_{10}]$;

③ 7 top observables in $\boldsymbol{X}^*$, $[x_5 \quad x_8 \quad x_7 \quad x_1 \quad x_{10} \quad x_4 \quad x_3]$;

④ 5 bottom observables in $\boldsymbol{X}^*$, $[x_4 \quad x_3 \quad x_9 \quad x_2 \quad x_6]$.

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 $\boldsymbol{X}^*$. 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 Tab. 3. 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.

Tab. 3 Estimated parameters with different observables

	$\hat{\theta}_5$	$\hat{\theta}_{12}$	$\hat{\theta}_{13}$	$\hat{\theta}_{16}$	$\hat{\theta}_{18}$
①	1.181	0.955	0.016 2	0.236 1	0.017 4
②	1.209	0.978	0.016 6	0.241 9	0.017 8
③	1.209	0.978	0.016 6	0.242 8	0.017 8
④	1.158	0.936	0.015 9	0.231 6	0.017 0

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.

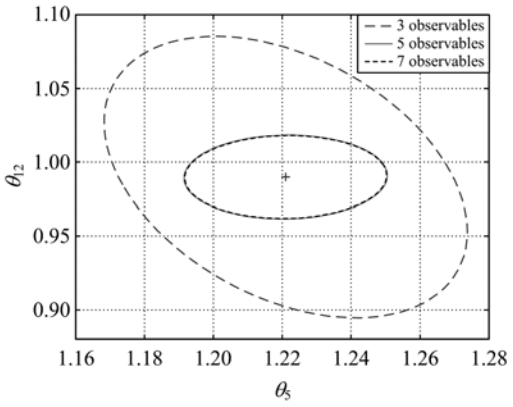

Fig. 1　Confidence interval of parameters θ_5 and θ_{12}

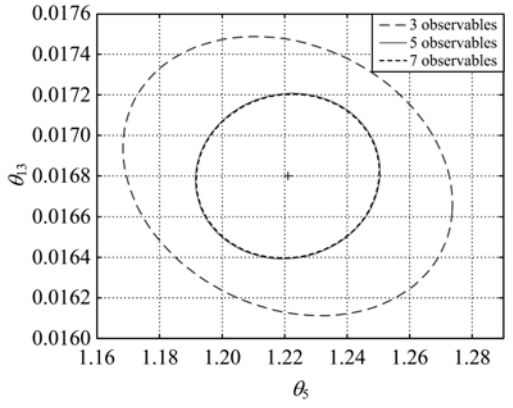

Fig. 2　Confidence interval of parameters θ_5 and θ_{13}

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

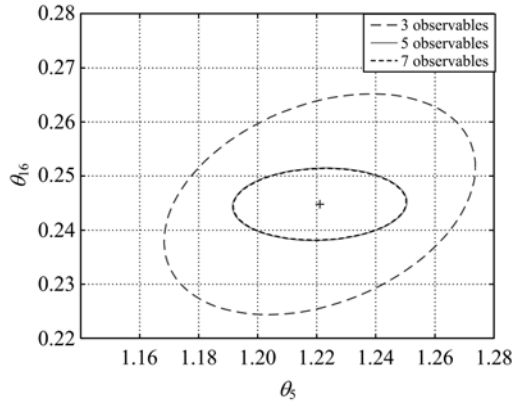

Fig. 3　Confidence interval of parameters θ_5 and θ_{16}

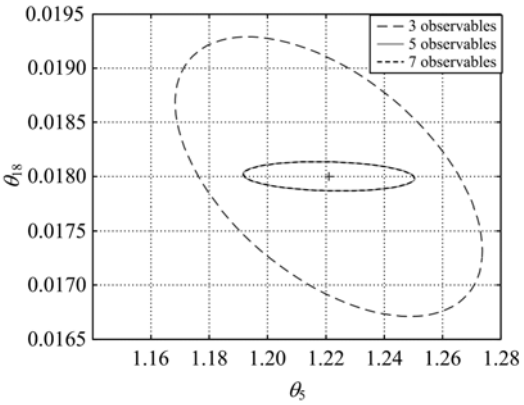

Fig. 4　Confidence interval of parameters θ_5 and θ_{18}

indistinguishable in Fig. 1 to Fig. 4. This result suggests that when the number of measurement variables used is less than the minimum 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 Tab. 3.

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 $\boldsymbol{X}^*$) is much smaller than the one using 5 non-optimal observables (bottom 5 states in $\boldsymbol{X}^*$).

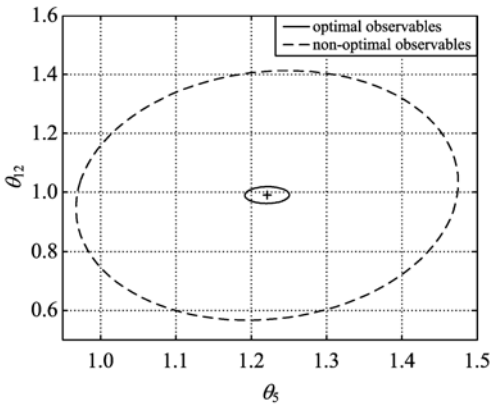

Fig. 5 Comparison of confidence interval of parameters θ_5 and θ_{12} w. r. t. the optimal and non-optimal observables

The former has a smaller parameter estimation error owing to the fact that the selected measurement set contains more information about the unknown parameters.

4 Conclusion

In this paper, the measurement set selection problem was discussed where the number of measurement variables and the priority of observables can be determined through matrix theory and model-based OED. In the studied example, it was 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.

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

References

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

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

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

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

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

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

[7] Franceschini G, Macchietto S. Model-based design of experiments for parameter precision: State of the art [J]. Chemical Engineering Science, 2008, 63(19): 4 846-4 872.

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

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

[10] He F, Brown M, Yue H. Maximin and Bayesian robust experimental design for measurement set selection in modelling biochemical regulatory systems [J]. International Journal of Robust and Nonlinear Control, 2010,24(6): 1 059-1 078.

[11] Boyd S, Vandenberghe L. Convex Optimization[M]. Cambridge: Cambridge University Press, 2004

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

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

[14] Hoffmann A, Levchenko A, Scott M L, et al. The IκB-NF-κB signaling module: Temporal control and selective gene activation[J]. Science, 2002,298:1 241-1 245.

[15] Nelson D E, Ihekwaba A E C, Elliott M, et al. Oscillations in NF-κB signaling control the dynamics of gene expression[J]. Science, 2004,306: 704-708.

[16] Lipniacki T, Paszek P, Brasier A R, et al. Mathematical model of NF-κB regulatory module[J]. Journal of Theoretical Biology, 2004,228(2):195-215.

[17] Ashall L, Horton C A, Nelson D E, et al. Pulsatile stimulation determines timing and specificity of NF-κB-dependent transcription[J]. Science, 2009,324(5924): 242-246.

[18] Jin Y S, Yue H, Brown M, et al. Improving data fitting of a signal transduction model by global sensitivity analysis [C]//Proc. American Control Conference, New York,USA, 2007: 2 708-2 713.