Optimal Experimental Design and Its Applications to Biochemical Engineering Systems

Hui Yu

Department of Electronic & Electrical Engineering

University of Strathclyde

This dissertation is submitted for the degree of

Doctor of Philosophy

July 2018

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Hui Yu

July 2018

Acknowledgements

I would like to express my deepest thanks and sincere gratitude to my supervisor, Dr. Hong Yue, for her all along support and excellent supervision. Her enthusiasm, inspiration, and rigorous attitude on research work profoundly influence me. Her thoughtful guidance, unlimited assistant, and critical remarks made this thesis possible. I really feel so lucky to have her as my supervisor.

I would like to express my deep gratitude to my second supervisor, Professor Peter Halling, for his knowledge, direction and perceptiveness. His noteworthy comments and constructive criticism on my work have been invaluable. I have benefited from his fascinating ideas on the research subject at every meeting with him.

I would also like to thank Dr. Jason Price and his colleagues in the Denmark University of Technology, for their valuable support of model information and useful discussions on my research work, and the cooperation between the two research groups.

Further sincere thanks to all my colleagues and friends at Wind Energy & Control Centre, The University of Strathclyde, for their encouragement and useful discussions on my research. The nice time we spend together would be memorable for me forever.

In the end, my special thanks to my parents, Aiying Wang and Yongan Yu, and my wife, Fang Ye, for their understanding, endless patience and constant support.

Abstract

This work is motivated by challenges in data-based modelling of complex systems due to limited information of sparse and noisy experimental data. Optimal experimental design (OED) techniques, which aim at devising necessary experiments to generate informative measurement data to facilitate model identification, have been investigated comprehensively. The limitations of existing experimental design approaches have been extensively discussed, based on which advanced experimental design methods and efficient numerical strategies have been developed for improved solutions. Two case study biochemical systems have been used through the research investigation, one is an enzyme reaction system, the other one is a lab-scale enzymatic biodiesel production system. The main contributions of this PhD work can be summarised as follows:

- Single objective experimental designs by considering one type of design factors, i.e. input intensity, measurement set selection, sampling profile design, respectively, has been formulated and numerical strategies to solve these optimisation problems have been described in detail. Implementations of these design methods to biochemical systems have demonstrated its efficiency in reducing parameter estimation errors.

- A new OED strategy has been proposed to cope with OED problems including multiple design factors in one optimisation framework. An iterative two-layer design structure is developed. In the lower layer for observation design, the sampling profile and the measurement set selection are combined and formulated as a single integrated

observation design problem, which is relaxed to a convex optimization problem that can be solved with a local method. Thus the measurement set selection and the sampling profile can be determined simultaneously. In the upper layer for input design, the optimisation of input intensities is obtained through stochastic global searching. In this way, the multi-factor optimisation problem is solved through the integration of a stochastic method, for the upper layer, and a deterministic method, for the lower layer.

- A new enzyme reaction model has been established which represents a typical class of enzymatic kinetically controlled synthesis process. This model contains important kinetic reaction features, moderate complexity, and complete model information. It can be used as a benchmark problem for development and comparison of OED algorithms. Systematic analysis has been performed in order to examine the system behaviours, and the dependence on model parameters, initial operation conditions. Structural identifiability and practical identifiability of this system have been analysed and identifiable parameters determined. The design of experiment for the enzyme reaction system by considering different types of design variables have been investigated. The parameter estimation precision can be improved significantly by using the proposed OED techniques, compared to the non-designed condition.

- The OED techniques are numerically investigated based on a lab-scale biodiesel production process with real experimental data through research collaboration with DTU in Denmark. The OED applications on this real system model allow to examine the effectiveness and efficiency of those new proposed OED methods. The measurement set selection and the sampling design of this system are developed which provide detailed instructions on how to improve experiments through OED. Also, the sensitivity analysis and parameter identifiability analysis are conducted; and their impacts to experimental design are clearly identified.

Table of contents

List of figures **xiii**

List of tables **xvii**

Nomenclature **xxi**

1 Introduction **1**

1.1 Overview 1

1.2 Aims and objectives 5

1.3 Novel contributions 6

1.4 Roadmap through this dissertation 8

1.5 Publications 10

2 Preliminaries to Data-based Modelling and Analysis **13**

2.1 Modelling procedure 14

2.2 Mathematical model representation 18

2.3 Parameter estimation 19

2.3.1 Parameter estimation methods 20

2.3.2 Assessing parameter estimation quality 23

2.4 Parameter sensitivity analysis 25

2.4.1 Local sensitivity analysis 26

2.4.2 Global sensitivity analysis 30

2.5 Identifiability analysis 31

2.5.1 Structural identifiability 31

2.5.2 Practical identifiability 32

2.6 Summary 38

3 Fundamentals and Applications in Optimal Experimental Design 39

3.1 Introduction 40

3.2 Brief review of statistical experimental design 41

3.3 Model-based experimental design 42

3.3.1 Experimental design for model discrimination 43

3.3.2 Experimental design for parameter estimation 48

3.3.3 Experimental design for input and observation factors 54

3.4 Numerical optimisation algorithms for solving OED problems 59

3.5 Summary 61

4 Dynamic Modelling of Two Biochemical Systems and Analysis 63

4.1 Establishment of a benchmark model - Enzyme reaction system 64

4.2 System analysis for enzyme reaction system 71

4.2.1 Equilibrium state analysis 71

4.2.2 Preliminary analysis for process optimisation 73

4.2.3 Local sensitivity analysis 74

4.3 Identifiability analysis for enzyme reaction system 87

4.3.1 Structural identifiability analysis 87

4.3.2 Practical identifiability analysis 89

4.4 Enzymatic biodiesel production system 96

4.5 Preliminary analysis to biodiesel production system 104

4.5.1 Local sensitivity analysis 104

4.5.2 Parameter identifiability analysis 106

4.6 Summary 111

5 Single Factor Experimental Design for Biochemical Systems 115

5.1 OED formulations for three individual design factors 116

5.2 Individual factor experimental design for enzyme reaction system 121

5.2.1 Optimal input design 121

5.2.2 Optimal sampling time design 135

5.2.3 Optimal Measurement set design 141

5.3 Individual factor experimental design for enzymatic biodiesel production system 143

5.3.1 Optimal sampling time design 145

5.3.2 Optimal Measurement set design 146

5.4 Summary 148

6 Comprehensive Experimental Design - A Two-layer Iterative Strategy 151

6.1 Integrated observation design of measurement set and sampling profile . . . 152

6.2 Iterative double layer design of observation and input 155

6.3 Case study 1: Comprehensive design for enzyme reaction system 160

6.3.1 Integrated observation design 160

6.3.2 Iterative two-layer design of observation and input 165

6.4 Case study 2: Integrated observation design for enzymatic biodiesel production system 166

6.5 Summary 171

7 Conclusions and Future Perspectives 173

7.1 Conclusions 173

7.2 Open challenges and future perspectives 176

References 179

Appendix A Identifiability analysis methods 193

Appendix B Preliminary analysis on multi-objective OED 199

Appendix C Preliminary analysis on robust experimental design 207

List of figures

Fig. 2.1 Model building and validation procedure based on model-based experimental design 15

Fig. 3.1 Model discrimination procedure based on OED 44

Fig. 3.2 Parameter identification process based on OED 49

Fig. 3.3 Influence of FIM maximisation on the parameter estimation cost function and confidence intervals for a single parameter estimation problem 52

Fig. 3.4 Geometrical interpretation of the standard criteria for OED 54

Fig. 3.5 An illustration of the CVP technique for the piecewise constant case over four intervals 56

Fig. 4.1 Reaction mechanisms of enzyme reaction system 65

Fig. 4.2 Time profiles for concentration of all state variables in enzyme reaction system 72

Fig. 4.3 Relationship between Q_{max} and initial input conditions 75

Fig. 4.4 Absolute parameter sensitivities to 10 state variables in the enzyme reaction system 77

Fig. 4.5 Relative sensitivities to state variable Q, E^*, E, and P by using log function 79

Fig. 4.6 Relative sensitivities to state variables and the corresponding parameter importance ranking 84

Fig. 4.7 Parameter ranking based on relative sensitivities for all state variables . 86

Fig. 4.8 Identifiability tableaus based on generating series approach 90

Fig. 4.9 Parameter importance ranking by orthogonalised sensitivity analysis & parameter correlations 93

Fig. 4.10 Collinearity index for all possible parameter subsets 95

Fig. 4.11 Identifiable parameter subsets for the enzyme reaction system model . 95

Fig. 4.12 Flowchart of the whole enzymatic biodiesel production process 96

Fig. 4.13 Enzyme kinetic transesterification process 99

Fig. 4.14 Time profile of 5 measurable state variables in enzymatic biodiesel production system 105

Fig. 4.15 Relative parametric sensitivities to state variable *BD* in enzymatic biodiesel production system 106

Fig. 4.16 Parameter rankings based on LSA for 5 measurable state variables in the biodiesel production system model 107

Fig. 4.17 Relative sensitivities to measurable state variables T and FFA 109

Fig. 4.18 The collinearity index values with different parameter subsets for the enzymatic biodiesel production system 110

Fig. 4.19 Parameter ranking based on the orthogonal backward elimination method for the enzymatic biodiesel production system 112

Fig. 5.1 Single input design of E_0, S_0, and N_0 with exhaustive searching method 124

Fig. 5.2 Comparison of different design criteria for single input design 129

Fig. 5.3 Comparison of parameter pair CIs for different input design factors . . 130

Fig. 5.4 Comparison of parameter pair correlations under different experiment conditions 131

Fig. 5.5 Comparison of CIs with local and global optimisation methods 134

Fig. 5.6 Optimal sampling points and relative sensitivities for (a) S, (b) Q, (c) E and (d) E^* 138

Fig. 5.7 CIs comparison of kinetic parameters under non-design sampling and the sampling schedule from D-optimal design 139

Fig. 5.8 Comparison of optimal sampling points with and without replicates . . 140

Fig. 5.9 Comparison of D-values from different sampling schedule 141

Fig. 5.10 Comparison of CIs of model parameters with different measurement sets 144

Fig. 5.11 CIs under D-optimal and emprical sampling strategies 147

Fig. 5.12 Comparison of CIs for k_2, k_6, k_8, and k_9 under different strategies of measurement sets 149

Fig. 6.1 Iterative double-layer design for both input and observation factors . . 158

Fig. 6.2 Sampling profiles of S and Q with different strategies 163

Fig. 6.3 Comparison of CIs for kinetic parameters under different observation design strategies 164

Fig. 6.4 Comparison of parameter uncertainties with different numerical strategies 167

Fig. 6.5 Sampling time points on the selected state variables 168

Fig. 6.6 Comparison of CIs for different strategies 170

List of tables

Table 4.1 Kinetic mechanism for the enzyme reaction system 66

Table 4.2 A list of initial values for state variables in the enzyme reaction system 70

Table 4.3 A list of nominal values for kinetic parameters in the enzyme reaction system 70

Table 4.4 Parameter pair correlations for the enzyme reaction system model . . 85

Table 4.5 Structural identifiability analysis with different number of measurable state variables by using generating series approach combined with identifiability tableau 91

Table 4.6 Parameter ranking based on the integrated effect on model outputs with PCA 92

Table 4.7 Kinetic mechanism for the enzyme reaction system 101

Table 4.8 Preliminary experiments for data fitting and validation 102

Table 4.9 Nominal parameter values estimated from measurement data following experiments 1 to 8 103

Table 4.10 Parameter pair correlations for the enzymatic biodiesel production system108

Table 4.11 Parameter importance ranking orders for the enzymatic biodiesel production system by using LSA and orthogonalised LSA 111

Table 5.1 Single input design results by using exhaustive search method 122

Table 5.2 Single input design results for enzyme reaction system by using active-set method 126

Table 5.3 Number of function evaluations required under two numerical strategies 127

Table 5.4 Uncertainties of selected kinetic parameters from A-, E-, ME-optimal experimental design for E_0 128

Table 5.5 Multi-input design results by using active-set method (local) 133

Table 5.6 Multi-input design results by using DE method (global) 133

Table 5.7 CIs of selected kinetic parameters under different experimental conditions by multiple input design 135

Table 5.8 Optimal sampling schedules from three searching methods and their corresponding D-values 136

Table 5.9 Optimal sampling time points and number of replicates in replicated experiments 137

Table 5.10 Measurement set design by considering five measurable state variables 142

Table 5.11 Measurement set design by considering ten state variables 142

Table 5.12 Sampling profiles under OED and empirical sampling strategies . . . 146

Table 5.13 Weighting coefficients for measurable state variables by measurement set selection design 146

Table 6.1 Three observation design strategies 155

Table 6.2 Observation design results for enzyme reaction system 161

Table 6.3 Input and observation experimental design results with different numerical strategies 165

Table 6.4 Design results of different OED strategies 169

Nomenclature

Roman Symbols

$\bar{s}_{i,j}$	scaled sensitivity of the i-th state variable over the j-th parameter
D	divergence of model predictions between different models
$\mathbf{F}$	parametric Jacobian matrix
$\mathbf{f}$	a set of state transition functions
$\mathbf{G}$	derivatives of model outputs w.r.t. parameters
g_i	derivative of the model equation w.r.t. the i-th input variable
$\mathbf{H}$	Hessian matrix
$\mathbf{h}$	a set of measurement functions
$\mathbf{J}$	Jacobian matrix
k_i	the i-th kinetic parameter
m	number of outputs that can be measured
n	number of state variables
N	number of sampling time points

p	number of parameters
$\mathbf{Q}$	measurement error covariance matrix
R	residual of parameter estimation
$\mathbb{R}^i$	the i-dimensional real number set
$\mathbf{S}$	parameter sensitivity matrix
$\mathbf{S}_0$	initial condition fo local parameter sensitivity matrix
$s_{i,j}$	absolute sensitivity of the i-th state variable over the j-th parameter
t	time in a continuous model
$\mathbf{T}_{sp}$	sampling design factors
$\mathbf{V}$	parameter estimation error covariance matrix
$\mathbf{W}$	weighting matrix
$\mathbf{X}$	vector of state variables
$\mathbf{X}_0$	initial condition of state variables
x_i	the i-th state variable
$\mathbf{Y}$	vector of measured model outputs
$\hat{\mathbf{Y}}$	vector of model outputs predictions
y_i	the i-th model output
$\hat{y}_i$	the i-th model output prediction
$\mathbf{z}$	measurement set design factors

Greek Symbols

χ^2	chi-square distribution
δ_i	CIs of the *i*-th model parameter
δ_j^{msqr}	mean squared root of absolute local sensitivites of *j*-th parameter to model outputs
γ	collinearity index
λ_{min}	minimal eigenvalue of a matrix
λ_{max}	maximal eigenvalue of a matrix
$\boldsymbol{\phi}$	vector of experimental design factors
σ_i^2	measurement error variance of y_i
$\boldsymbol{\theta}$	vector of model parameters
Θ	admissible range of parameters
$\hat{\boldsymbol{\theta}}$	estimated model parameters
θ_i	the *i*-th model parameter
υ	objective function of OED
ς	input design factors
$\boldsymbol{\xi}$	vector of measurement errors

Superscripts

T	transpose of a vector

Other Symbols

$\arg\min f(x)$ value of argument x that minimise $f(x)$

Acronyms / Abbreviations

AIM	analytically integrated Magnus method
CI	confidence intervals
CRLB	Cramer-Rao lower bound
CVP	control vector parametrisation
DDM	direct differential method
FDM	finite difference method
FIM	Fisher information matrix
GAs	genetic algorithm
GFM	green function method
GSA	Global sensitivity analysis
i.i.d.	identically and independently distributed
LSA	local sensitivity analysis
LSE	least square estimation
MLE	maximum likelihood estimation
NLP	non-linear programming
$NP-hard$	non-deterministic polynomial-time hard
ODEs	ordinary differential equations

OED	Optimal experimental design
OFAT	one-factor-at-a-time
PDEs	partial differential equations
PSO	Particle swarm optimisation
RED	Robust experimental design

Chapter 1

Introduction

1.1 Overview

Mathematical models are widely used in chemical and biological systems since the mathematical representation allows us to reproduce real dynamic processes in a simulation environment (Peleg et al., 2002; van Riel, 2006a). These models can be used to explore the underlying nature of specific reactions, to better understand the dynamics of individual components and their interactions, to predict and control the future behaviour of systems. In the past few decades, numerous methods of quantitative modelling for chemical and biochemical systems have been proposed and well applied in sciences and engineering (Asprey and Macchietto, 2000; Phair, 1997; Villaverde et al., 2014).

During the modelling process, one intends to link inputs, which are disturbances and manipulations, to outputs that are of interest to the users following certain physical laws. The relationship between inputs and outputs specifies the basic structure of the model, which can then be described by models that consist of variables and parameters. These model parameters are usually unknown and need to be estimated from real measurement data, the sampling of which can greatly affect the quality of parameter estimation (Ashyraliyev et al., 2009; Sun et al., 2012). The problem of parameter identification itself is a non-trivial task especially for complex systems with unknown dynamic nature. It becomes even more difficult if the experimental data is less informative, sparse and contaminated with noise (Feng and Rabitz, 2004; Ljung and Wills, 2010; Vanlier et al., 2013). It is therefore very important to plan the experiment in priori to make sure the generated measurement data are informative and relevant to questions to be addressed. On the other hand, when prior information of a system is insufficient, some hypotheses need to be postulated on the behaviour of the process. Each assumption can lead to a specific model structure. In this case, a set of models will be proposed, which can fit preliminary experimental data and describe the process behaviour. It is necessary to distinguish all the candidates and find the one that can best describe the process and make the most accurate prediction of system behaviour.

The most efficient way of model discrimination is perhaps to find a set of measurement data that can maximise the divergence among those proposed models so as to identify the best fitted model structure (Kremling et al., 2004; Wiens, 2009). The experimental data are thus of particular importance in both parameter estimation and model discrimination. A well-planned experiment can lead to more informative data that facilitates systems identification; while a poor-planned experiment may probably result in limited data information and a waste of resources, even leading to erroneous conclusions in some cases. This motivates research work on the model-based optimal experimental design (OED) for chemical and biochemical systems (Jaqaman and Danuser, 2006; Kreutz and Timmer, 2009; Pronzato, 2008).

The purpose of OED is to devise necessary experiments for systems to be modelled in order to obtain the most informative measurement data, which can then be used to estimate unknown model parameters with the best statistical quality or facilitate model discrimination (Atkinson, 1996; Franceschini and Macchietto, 2008). This thesis work is mainly focused on the first part, which is referred to as the model-based OED for parameter estimation.

When a model structure is established, the central task is to estimate unknown model parameters from experimental data. Intuitive experimentation or that from expert experiences may result in inadequate data information. Furthermore, for large complex systems with high non-linearity and poorly understood dynamics, measurements usually cannot be taken for all state variables of the system, and the measured data is subject to substantial experimental noise, not to mention the intensive experiment cost and large amount of time required for conducting the experiments (Huang and Wu, 2008; Omony et al., 2012; Pagendam and Pollett, 2013). Therefore, the OED technique becomes essential in the selection of experiment set so that the most valuable data can be generated to increase parameter estimation precision and reduce the experimental efforts. For this purpose, it is necessary to incorporate OED into parameter estimation procedures (Hagen et al., 2013; Silvey, 2013).

The experiment set can be grouped into: *input manipulations* and *observation strategies*. Manipulation of inputs will drive the dynamic system and change the dynamic responses of the system. Typical inputs include initial concentrations of reactants and external time variant or invariant input variables. The design of observations is focused on the choice of measurements, that is, to select a subset of informative data from overall possible experimental data. In the design of observation strategies under given input conditions, the system response will not change. In the past few decades, a lot of interesting work has been published on the development of OED methods, such as input design (Baltes et al., 1994; Chianeh et al., 2011; Chung et al., 2000; Dirion et al., 2008; Kremling et al., 2004; Lindner and Hitzmann, 2006; Zak et al., 2003), sampling time design (Alaña and Theodoropoulos, 2012; Asyali, 2010; Ataíde and Hitzmann, 2009; de Brauwere et al., 2009; Gil et al., 2014; Paquet-Durand et al., 2015; Zhu and Stein, 2006), measurement set selection (Bansal et al., 2013; Brown et al., 2008; He et al., 2010; Yu et al., 2015), to name a few. However, most of these works are only focused on a single design factor, e.g., the input design of a mitogen-activated protein kinase (MAPK) signalling pathway model (Faller et al., 2003). The integration of multiple design factors is more useful in practice as it will produce more informative data than by looking at a single design factor each time because the design factors may interact with each other. When multiple design factors are considered in an OED, it will normally make a large non-linear dynamic programming problem (Balsa-Canto et al., 2008; Bauer et al., 2000). The development of efficient numerical strategies and algorithms are required to solve such complex dynamic optimisation problems.

In the context of model-based OED, the Fisher information matrix (FIM) is probably the most widely used method to quantify the information of experimental data. The FIM is a matrix that contains measurement error information as well as information about model parameter effects on model outputs. It is not easy to compare data information from different experimental strategies through the direct comparison of information matrices. Various scalar functions of FIM have been proposed as metrics to measure the data information. These scalar design criteria, named 'alphabetic' design criteria, are concerned with different properties of the confidence region of parameter estimates (Anderson-Cook, 2007; Hosten, 1974; Ljung, 1998). Each design criterion has its own pros and cons, and none of them suits all cases. For example, the D-optimal design focuses on the reduction of volume of the confidence region of parameter estimation thus reduces the overall parameter estimation errors, but it does not consider any parameter correlations. The modified E-optimal design, in contrast to the D-optimal design, mainly focuses on the reduction of parameter correlations. Therefore, it is necessary to try these scalar design criteria in order to support more accurate and independent parameter estimates.

In the OED framework, the FIM is constructed from local parametric sensitivities, the calculation of which depends on the nominal values of model parameters. However, the values of model parameters in biochemical systems are usually very uncertain a priori. One option to deal with this problem is to use the sequential experimental design procedure (Buzzi-Ferraris and Forzatti, 1983; Hagen et al., 2013; Hering and Šimandl, 2010; Schwaab et al., 2006). The parameter identification and OED are implemented iteratively until satisfactory parameter estimation is achieved. This iterative procedure normally involves a large number of experiments, which are cost intensive and time-consuming. Thus robust experimental design (RED) techniques are required to tackle the design problems under large parameter uncertainties.

Another interesting aspect is the selection of important parameters for the dynamic systems under consideration. In most chemical systems, the effects of model parameters on the system outputs are different from one to another and it is usually difficult to estimate all model parameters with satisfactory quality (McLean and McAuley, 2012). It is essential to determine whether the lack of identifiability is caused by the model structure or by the limited experimental data. It is crucial to identify those important parameters that have significant impact on the model outputs. Parameter estimation can then be focused on those important parameters, while the other parameters which have little effects on system outputs can be set at fixed values or even removed from the model (Chu and Hahn, 2007; Farina et al., 2006).

1.2 Aims and objectives

The aim of this PhD work is to develop efficient and effective experimental design approaches so that the parameter estimation quality can be improved using the data from the designed experiment. These methods are developed for chemical, biochemical and wider systems with high-dimensional and poorly known non-linear dynamics.

The main research objectives are listed below.

1. To construct experimental design formulations by considering key design factors, i.e., input design, sampling time design and measurement set selection, individually, and investigate numerical optimisation strategies to solve these OED problems.

2. To develop efficient optimisation procedures for the comprehensive design of multiple experimental factors such as initial condition, sampling time strategy, measurement set selection and others, in an integrated framework rather than designing each individually.

3. To implement OED techniques to typical biochemical systems and examine their efficiency and effectiveness in parameter estimation. Investigate parameter sensitivity and parameter subset selection approaches. Identify their roles and connections with model-based experimental design.

1.3 Novel contributions

The main contributions of this PhD work can be summarised as follows.

- Develop OED formulations and numerical optimisations by considering input intensities, sampling time profiles, and measurement sets, individually. This is referred to as single factor experimental design.

  The OED problems for various design factors, such as input intensity design, sampling time profile design and measurement set selection, have been formulated as proper optimisation problems. Various numerical strategies have been developed and examined on their ability to find global optima and numerical efficiency. These OED methods are applied to an enzyme reaction model in order to provide guidance on how to use model-based experimental design approaches to facilitate model identification, particularly for parameter estimation.

- Propose a novel numerical strategy to integrate multiple experimental design factors into one OED framework

  The integrated observation design that determines both measurement set selection and sampling time scheduling, simultaneously, has been proposed. By approximating available sampling points a priori, the problem formulation for sampling time design can be expressed in a similar form as measurement set selection design. Therefore these two design tasks can be combined together as a single objective optimisation problem, which is further relaxed to a convex optimisation problem that can be conveniently solved using local optimisation methods. Furthermore, an iterative two-layer numerical strategy has been developed to deal with comprehensive OED taking into account input and observation variables together. This new optimisation strategy intends to obtain optimal results for all experimental conditions in one optimisation framework. The input design that is formulated as a non-convex optimisation problem is solved by modern heuristic algorithm, PSO method, in the upper layer. The integrated observation design which can be relaxed into convex optimisation problem is solved by a local optimisation method the Powell's method, in the lower layer. The two layers are iterated until the threshold or the limit of iteration number is reached.

- Establish a new enzyme reaction model that can be used for benchmark OED studies.

  A new enzyme reaction system model is established which can represent a typical class of enzymatic processes. Systematic analysis for this enzyme reaction model has been performed to examine the model characteristics. The design of experiments for this system with the proposed OED methods have been comprehensively investigated. A number of enzymatic kinetically controlled synthesis processes follow the similar reaction scheme. Model identification problems are often encountered in this kind of systems due to sparse and noisy experimental data, as well as unknown complex dynamic nature. It is therefore of particular interest to investigate this enzyme reaction model as it can provide a comprehensive understanding on model identification of a wide group of enzyme process models.

- Investigate OED methods to a lab-scale biodiesel production system with real data.

A lab-scale enzymatic biodiesel production system model, which contains seventeen state variables and twenty kinetic parameters, has been employed as a basis for OED investigation with support from collaborator at DTU, Denmark. This model contains typical kinetic features, moderate complexity and real experimental data. The OED investigation for this model examines the effectiveness and efficiency of the new proposed OED methods. It can also provide guidance on the implementation of OED techniques to real chemical or biochemical systems.

In addition, preliminary investigation on the robust sampling time design for the enzyme reaction system has been done which is provided in Appendix B. Also, the integration of process optimisation and OED is also another interesting research aspect. Preliminary analysis has been taken which is given in Appendix C.

1.4 Roadmap through this dissertation

In this PhD work, the model-based OED techniques have been investigated with applications to two biochemical systems. The main purpose is to develop new approaches to generate informative measurement data which can facilitate parameter estimation during model building process. To support efficient OED, local sensitivity analysis (LSA), model identifiability analysis and parameter estimation have also been studied. An overview of each chapter is given as follows.

Chapter 2 **Preliminaries to Data-based Modelling and Analysis** provides preliminaries of mathematical modelling for biochemical systems and the challenges during the model building and validation process. A focus is on understanding the roles and importance of OED in systems identification. Relevant topics such as parameter estimation, sensitivity analysis and identifiability analysis have been discussed. Numerical methods on conducting these analyses and their links to OED techniques are explained.

Chapter 3 **Fundamentals and Applications in Optimal Experimental Design** reviews existing OED techniques. Firstly, a brief review of earlier statistical experimental design methods is given. The limitations of this kind of methods are discussed, which supports the rationale of model-based OED approaches. Then recent development and applications of model-based OED for model discrimination and parameter estimation are reviewed. Finally, typical numerical approaches on solving the OED problems are introduced.

Chapter 4 **Dynamic Modelling of Two Biochemical Systems and Analysis** provides modelling of two biochemical systems. The first one is an enzyme kinetically controlled synthesis system which can be used as a benchmark system for OED investigation. The second one is a lab scale enzymatic biodiesel production system with real experimental data. These two models are used as exemplars to illustrate the function and efficiency of the proposed OED methods throughout the whole thesis. Dynamic and steady-state behaviours of the two systems have been examined. The LSA and identifiability analysis of these two biochemical models are discussed in detail, from which key parameters for the outputs of interest are determined. Those important parameters are the focus in OED.

Chapter 5 **Single Factor Experimental Design for Biochemical Systems** investigates the OED to biochemical systems when a single design factor is considered. The OED formulations and the numerical solution strategies are investigated. The functionality of OED techniques for improving parameter estimation accuracy is testified through the two case studies. Overall, this chapter provides guidance on how to systematically design experiments for model development.

Chapter 6 **Comprehensive Experimental Design - A Two-layer Iterative Strategy** investigates OED problems when multiple design factors are considered. A new double-layer optimisation algorithm has been proposed which combines input and observation designs together. The input design that will change system dynamic responses is solved in the upper layer with a global optimisation algorithm. In the observation design, the measurement set selection and the sampling time profile design are combined into one single objective OED, which can be transferred to a convex optimisation problem and solved in the lower layer by a local optimisation method. The whole process is an iterative session between the two layers. The efficiency of the proposed method has been examined through the enzyme reaction system.

Finally, Chapter 7 **Conclusions and Future Perspectives** presents the main research contributions from this thesis and discusses potential future researches for OED.

In addition, some preliminary results on the RED for sampling time profile, and the integration of process optimisation and OED are given in Appendix B and C, respectively.

1.5 Publications

- Yue, Hong, Peter Halling, and Hui Yu. "Model development and optimal experimental design of a kinetically controlled synthesis system." The 12th IFAC Symposium on Computer Applications in Biotechnology (CAB), Mumbai, India. December 16-18, 2013, Volumes 46.31: 327-332. (Chapter 4)

- Yu, Hui, Hong Yue, and Peter Halling. "Optimal experimental design for an enzymatic biodiesel production system." The 9th IFAC Symposium on Advanced Control of Chemical Processes (ADCHEM), Whistler, British Columbia, Canada. June 07-10, 2015, Volumes 48.8: 1258-1263. (Chapter 5, case study 2 for the enzymatic biodiesel production system)

- Yu, Hui, Hong Yue, and Peter Halling. "A two-loop optimisation strategy for multi-objective optimal experimental design." The 11th IFAC Symposium on Dynamics and Control of Process and Bioprocess Systems (DYCOPS), Trondheim, Norway. June 06-08, 2016, Volumes 49.7: 803-808. (Chapter 6)

- Yu, Hui, Hening Yu, Hong Yue, and Jinglin Zhou. "Integrated time sampling design and measurement set selection for biochemical systems." The 22nd International Conference on Automation and Computing (ICAC), IEEE. Colchester, United Kingdom. September 07-08, 2016. (Chapter 5, case study 1 for the enzyme reaction system)

- Hui Yu, Hong Yue and Peter Halling. "Robust sampling time design for enzyme reaction system." The 10th International Symposium on Advanced Control of Chemical Processes (ADCHEM), Shenyang, China. July 25-27, 2018. (accepted) (Chapter 7)

- Yu, Hui, Hong Yue, and Peter Halling. "Comprehensive experimental design for chemical engineering processes: A two-layer iterative design approach." Chemical Engineering Science, May 25, 2018. (Chapters 4-6)

Chapter 2

Preliminaries to Data-based Modelling and Analysis

In this chapter, the general mathematical modelling and identification procedures of chemical and biochemical systems have been described. Problems and challenges in model identification are discussed and the importance of experimental design to model building is highlighted. In addition, the basic parameter estimation scheme, parametric sensitivity analysis and model identifiability analysis are introduced. They are fundamental to a good understanding of OED and also crucial to the development of OED methods. In Section 2.1 the general modelling procedure and challenges related to identifiability analysis and experimental design are discussed. Section 2.2 gives general mathematical modelling formalisms and specifically provides a detailed explanation on nonlinear dynamic model representation in the form of ordinary differential equations (ODEs). Section 2.3 reviews the most widely used parameter estimation methods and their corresponding adequacy test tools. Section 2.4 discusses parametric sensitivity analysis which plays a key role in parameter estimation, identifiability analysis and experimental design. Section 2.5 presents model identifiability analysis and discusses its role in experimental design.

2.1 Modelling procedure

A chemical or biochemical model normally includes two aspects: (i) the topology which indicates the interconnections between different reaction species; (ii) the dynamics which describes the nature and property of those interactions, usually determined by model parameters (Asprey and Macchietto, 2000). Biochemical modelling intends to identify the most suitable model structure and obtain the most accurate model parameter values (Alberton et al., 2011; Tommasi, 2009). The final model should have the ability to mimic the real processes with moderate model complexity. As in the first part of this PhD work, the overall framework of model building and validation procedure including OED is established. This framework is illustrated in Figure 2.1, which is expanded from the model building procedure in (Franceschini and Macchietto, 2008).

Basically the modelling procedure consists of four main steps: preliminary analysis, model discrimination, parameter estimation and model validation. Once one or several candidate models are proposed from prior knowledge, the next task is to investigate whether it is possible to obtain unique solutions for model parameters under the candidate model(s). If not, alternative models need to be proposed. For those structurally identifiable models, parameter sensitivity analysis and practical identifiability analysis can be performed to figure out crucial model parameters which have significant effects on model outputs. This will make further model calibration more focused on those key parameters, whereas model parameters which have little influence on model outputs can be kept at their nominal values or removed so as to reduce the model complexity. After this preliminary study, surviving models will be compared through the experimental data fitting process so that the most suitable model can be determined which makes the best model description for the studied system. Next the selected model will undergo further experimental data fitting process in order to improve the parameter estimation accuracy (this is also called model calibration or parameter estimation).

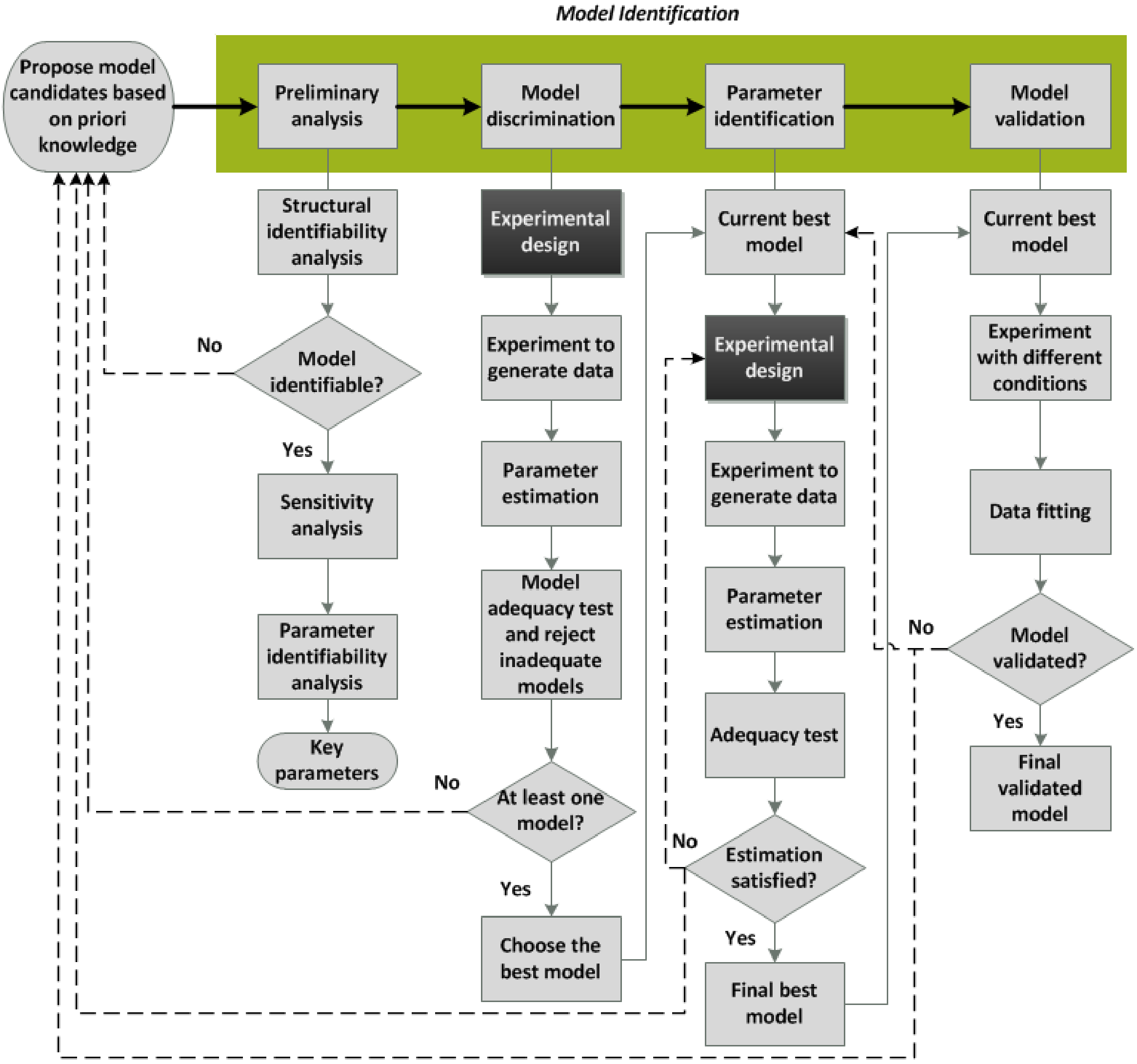


Fig. 2.1 Model building and validation procedure based on model-based experimental design

Finally, the calibrated model needs to be tested by data sets taken from various experimental conditions in order to validate the selected model. In Figure 2.1, those dashed lines imply that model building is an iterative process. For example, if the selected model structure is found to be not appropriate, it is necessary to propose new model(s) and repeat the modelling procedure until a final validated model is achieved.

Experiments in real processes are constrained by operation conditions and measurement. In many cases only a limited number of reactants in the biochemical network can be measured. Specific ways of input stimulation also confine the types of experimentation. The number of experiments could be limited due to the restrictions of budget and time resources. Also, the measurement data are inevitably contaminated with noise. All these constraints, combined with the complex non-linear dynamic nature of biochemical systems, make model identification a very challenging task (Chu et al., 2009; Thompson et al., 2009). It is essential to provide high quality and informative measurement data to facilitate model identification with reduced experimental cost. Therefore, the OED techniques play key roles in model building and validation process. The purpose of model-based OED is to devise necessary experiments to generate efficient measurement data for the identification of model structure and the parameter estimation with reduced experimental efforts.

Generally speaking, in the modelling of biochemical networks, the model framework is determined by the following considerations: the purpose of the modelling; key experimental factors for the questions to be addressed; the assumptions about the reactions and species involved; together with the relevant biochemical and physical laws. It is common to propose several model structures that may all represent the system behaviour, to some extent, and have similar predictive capabilities (Atkinson and Fedorov, 1975a; Box and Hill, 1967; López-Fidalgo et al., 2008). In order to identify the model with the most suitable structure, experiments should be designed in such a way that the divergence between different model structures is maximised. Once the structure of the model is determined, model parameters will need to be evaluated by comparing model predictions to available experimental data. For this purpose, it is essential to obtain quantitative and informative experimental data. However, performing time-course experiments for biochemical or biological systems, is usually expensive and time-consuming. OED techniques can help to generate the most informative data so as to increase parameter estimation precision without unnecessary experimental efforts. During past few decades the OED has attracted a lot of interest and has been successfully applied in a wide range of systems (Casey et al., 2007; Chianeh et al., 2011; Omony et al., 2012; Sjögren et al., 2011; Skanda and Lebiedz, 2010; Strigul et al., 2009; Versyck and Van Impe, 1998; Walter and Pronzato, 1990). The detailed review about experimental design will be given in Chapter 3.

Model identification and experimental design are still open and active research areas. Biochemical networks normally contain a series of parallel, consecutive and competitive reversible reactions, which may lead to strong correlations between model parameters. The influences of parameters on model outputs may not be at similar levels. In addition, kinetic parameters are usually very uncertain a priori. All these factors indicate that not all model parameters of a biochemical system can be assessed with unique and meaningful values. Several methods have been proposed based on parameter sensitivity analysis or FIM that can deal with parameter identifiability problem (Brun et al., 2001; Raue et al., 2009; Yao et al., 2003).

With regards to the OED problems, most published work only focuses on a single design factor, e.g. input design, sampling time design, or measurement set selection. Very few researchers considered the integration of multiple design factors in one OED framework (Lindner and Hitzmann, 2006). Considering the fact that experimental factors are often dependent on each other, it would be ideal to design the experiment that could take into account crucial experimental factors together. This should make the measurement data generated from the intended experiments contains more information compared to single factor design. An optimisation problem of OED that involves multiple design factors is usually an NP-hard problem which is difficult to solve (Bauer et al., 1999; Lohmann et al., 1992).

2.2 Mathematical model representation

Continuous deterministic dynamic models described by ODEs are the most prominent model descriptions in recent biochemical systems literature. Consider a general biochemical model with n state variables and p parameters, the state transition and output models can be written as follows:

$$\begin{aligned} \dot{\mathbf{X}}(t) &= \mathbf{f}(\mathbf{X}(t),\boldsymbol{\theta}),\ \mathbf{X}(t_0) = \mathbf{X}_0 \\ \mathbf{Y} &= \mathbf{h}(\mathbf{X}(t),\boldsymbol{\theta}) + \boldsymbol{\xi}(t) \end{aligned} \tag{2.1}$$

where $\mathbf{f}(\cdot)$ is a set of state transition functions of the system dynamics which are assumed to be continuous and first-order derivative; $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ denotes the vector of n state variables with initial condition $\mathbf{X}_0$; $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T \in \mathbb{R}^p$ is the vector of p model parameters; $\mathbf{Y} = [y_1, y_2, \ldots, y_m]^T \in \mathbb{R}^m$ is the measurement output vector with m ($m \leq n$) measurable variables; $\mathbf{h}(\cdot)$ is the measurement function, normally used for selecting which variables to be measured; $\boldsymbol{\xi}$ is the vector of measurement errors which can be classified into systematic errors and random errors. The experiments should be designed to eliminate the systematic errors. However, the random errors that disturb the observations always exist. Most often the measurement error is assumed to be a zero mean, Gaussian white noise. Two important assumptions are made for this ODEs representation. Firstly the process is assumed to be homogeneous and 'well-stirred'. Otherwise, spatial effect needs to be considered which should be represented by partial differential equations (PDEs). Secondly, the concentrations of all the reactants in the system are continuous functions of time.

A two-step Michaelis-Menten kinetic reaction model is given, as an example, to explain how to obtain the mathematical model from biochemical reactions. The chemical reactions are shown below:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} P$$

The first step with double arrow depicts reversible reactions of the binding of substrate S and enzyme E, and the decomposition of enzyme substrate complex ES, with rate constants k_1 and k_{-1} for forward and backward processes respectively. The second reaction is irreversible where enzyme substrate complex ES is converted to the product P with rate constant k_2. Following the mass action laws, the set of ODEs for state variables, E, S, ES and P, are constructed as follows:

$$\begin{aligned}
&\tfrac{dE}{dt} = -k_1 E \cdot S + k_{-1} ES + k_2 ES \\
&\tfrac{dS}{dt} = -k_1 E \cdot S + k_{-1} ES \\
&\tfrac{dES}{dt} = k_1 E \cdot S - k_{-1} ES - k_2 ES \\
&\tfrac{dP}{dt} = k_2 ES
\end{aligned}$$

This model is linear in parameters and non-linear (bilinear) in state variables. Similar model structures exist in a lot of biochemical systems, e.g. enzyme reaction systems and enzymatic biodiesel production systems that will be discussed later on in this thesis.

2.3 Parameter estimation

Once the model structure is decided, the primary task is to identify the model parameters (kinetic constants), which are often roughly known from literature or approximately estimated from preliminary experiments. The most widely used method to determine these parameter values is to fit measurable output variables from model simulations to real measurement data (Di Maggio et al., 2010; Peifer and Timmer, 2007). Such problems are called inverse problems or parameter estimation problems. In parameter estimation, researchers not only need to obtain the mathematical solution, but also need to justify whether those parameters are physically plausible, to examine the predictive ability of the model, and to check if it is consistent in other experimental conditions or with other constraints (Holford, 2005; Schittkowski, 2007). Issues related to parameter estimation include existence of solution(s), uniqueness of solution and computational stability. Due to the complex non-linear dynamic nature of biochemical systems and limited measurement data corrupted with noise, as well as a large number of parameters with correlations between each other, it is quite often that there is no solution or multiple solutions for parameter estimation (Esposito and Floudas, 2000; Yen et al., 1998). Global optimisation of this fitting problem to obtain precise parameter estimates is very difficult. Furthermore, in some cases the computational process for the inverse problem is extremely unstable where a small change of measurement data can make a tremendous change of model parameter values, which is referred to as ill-posed parameter estimation problem (Brown et al., 2010). Numerous efforts have been made on the development of parameter estimation techniques to surmount these problems (Hug et al., 2013; Moles et al., 2003). In this section, parameter estimation techniques are briefly reviewed with the focus on the least square parameter estimation (LSE) framework. The statistical analysis of LSE is also introduced.

2.3.1 Parameter estimation methods

Generally parameter estimation techniques can be grouped as classic approaches and Bayesian estimation methods. The solutions of parameter estimation from these two groups of methods are conceptually different. In classic approaches, the model parameters are assumed to be deterministic; while in a Bayesian estimator, the model parameters are assumed to be stochastic and take the form of a probability distribution. The main idea of Bayesian estimation is to calculate the posterior distribution of model parameters based on the prior parameter probability distribution and current data observation (Hug et al., 2013; Sun and Sun, 2015). Due to the cumbersome computational load and the requirement of prior knowledge about the probability distribution of model parameters, the application of Bayesian estimator lags far behind the theory. The most prevalent method of parameter estimation for biochemical networks is the (weighted) LSE, which is a special case of the maximum likelihood estimation (MLE) with the assumption of uncorrelated, normally distributed measurement noise. Mathematically, it is performed by optimising a scalar cost function $R(\boldsymbol{\theta})$ with respect to model parameters. The cost function is usually a weighted measure of the divergence of the predicted value of observables from the experimental data. Following the LSE framework, parameter estimation for system equations in (2.1) can be cast as a (large) non-linear dynamic programming problem subject to non-linear differential-algebraic constraints (Ciucci, 2013; Sun and Sun, 2015).

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}) \\
&= \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{l=1}^{N} \left(\mathbf{Y}(t_l) - \hat{\mathbf{Y}}\left(\hat{\boldsymbol{\theta}}, t_l\right)\right)^T \mathbf{Q}^{-1} \left(\mathbf{Y}(t_l) - \hat{\mathbf{Y}}\left(\hat{\boldsymbol{\theta}}, t_l\right)\right) \\
&= \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{m} \sum_{l=1}^{N} \frac{1}{\sigma_i^2} \left(y_i(t_l) - \hat{y}_i(\boldsymbol{\theta}, t_l)\right)^2
\end{aligned} \tag{2.2}$$

where $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are measured values and model prediction of the output vector at sampling times t_l $(l = 1, 2, \cdots, N)$, N is the total number of sampling data in time. $\mathbf{Q} \in \mathbb{R}^{m\times m}$ is the measurement error covariance matrix, the inverse of which is taken as a weighting matrix to state variables $\mathbf{Y}$. Assuming all observation variables can be measured independently and characterised by the variance of σ_j^2, the measurement error covariance matrix is written as $\boldsymbol{Q} = \mathrm{diag}[\sigma_1^2, \cdots, \sigma_m^2]$. σ_i^2 denotes the measurement error variance of the i-th state variable. Smaller measurement error variances will lead to larger weights to the corresponding measurement indicating more contribution of the measurement to the objective function. If the model is linear in parameters and the output function $\mathbf{Y}$ can be expressed as

$$\mathbf{Y}(\boldsymbol{\theta}) = \mathbf{G} \cdot \boldsymbol{\theta} \tag{2.3}$$

then the least square solution for $\boldsymbol{\theta}$ can be determined to be

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{GY} \tag{2.4}$$

For models that are non-linear in parameters, there is no general method to solve the parameter estimation problem. A common way is to linearise the model around the nominal parameter values and apply linear LSE method to solve the inverse problem.

There are several issues that need to be considered during parameter estimation. One issue is that the parameter estimation results are easily affected by the actual measurement data, especially by large outliers which often exist in real systems data sets (Jaqaman and Danuser, 2006). Therefore, experiments need to be carefully conducted to eliminate system errors. Robust parameter estimation techniques can be used, such as least median of squares (Massart et al., 1986; Mintz and Meer, 1991). Another issue is the instability of computational process for parameter estimation. In most biochemical networks, model parameters are usually highly correlated. One can either simplify the model to reduce the number of parameters or apply regularisation tools, such as the Tikhonv regularisation method (Sun and Sun, 2015), so that parameter values can be uniquely determined. For complex non-linear biochemical systems with noisy measurement data, several numerical strategies have been proposed for parameter estimation, such as multiple shooting methods (Peifer and Timmer, 2007), stochastic algorithm based on probabilistic algorithms (Catania and Paladino, 2009) and hybrid methods integrating global optimisation with gradient based optimisation (Balsa-Canto et al., 1998; Banga et al., 2005; Rodriguez-Fernandez et al., 2006).

2.3.2 Assessing parameter estimation quality

When the parameters are estimated, it is necessary to assess how precise the model parameters are and whether the model can explain the measurement data satisfactorily. The sum of the squares of the residuals, which is attributed to measurement noise and lack-of-fit of the model, contains useful information about the quality of model estimates (Di Maggio et al., 2010). As the measurement noise is assumed to be identically, independently distributed (*i.i.d.*) Gaussian white noise, the residual χ^2_{obs} follows a χ^2 (chi-square) distribution with $N-p$ degrees of freedom. This value is tested with reference to the standard χ^2 distribution value. A good fit of model to measurement data should be the one at which χ^2_{obs} is smaller than the standard χ^2 distribution value (Devore, 2015; Motulsky and Christopoulos, 2004). Note that a very small residual may indicate the over fit of model to measurement observation. When the standard deviations of measurement errors are not known, the variance of measurement errors needs to be determined via repeated estimation and in this case F-test is required to help assessing the quality of a model.

In addition to the χ^2 test or F-test, the significance of model parameters also needs to be evaluated. From (2.4) it is known that parameters estimated via linear regression are linear combinations of measurement data. The values of model parameters also follow normal distribution with the assumed statistical properties of measurement noise. To evaluate the statistical significance of parameter estimates, the student t-value and the joint confidence regions between parameters are two widely used measure. The student t-value is tested by comparing with the standard t-distribution value in order to assess the uncertainty for each parameter (Kolaczyk and Csárdi, 2014). Higher t-values indicate reliable estimates while lower values suggest larger uncertainties of the parameters. However, this method does not consider the correlations between model parameters, which may cause very low t-values. The parameter joint confidence interval (CI) is another useful metric to assess parameter estimation precisions. It is used to locate the parameter estimation confidence region in which it is expected that the true parameters lie. For example, a 95% confidence region means that the true parameters fall into the region with 95% probability. An exact confidence region can be determined based on the cost function in (2.3) (Motulsky and Christopoulos, 2004):

$$\left\{\boldsymbol{\theta} : R(\boldsymbol{\theta}) \leq \left(1 + \frac{p}{N-p} \times F_{p,N-p}^{1-\alpha}\right) \times R(\hat{\boldsymbol{\theta}})\right\} \tag{2.5}$$

where $F_{p,N-p}^{1-\alpha}$ is the upper α critical level F distribution with p and $N-p$ degrees of freedom. However, (2.5) is inconvenient for numerical implementations. Most often linear approximations are used to construct the confidence regions. For models that are linear in parameters, the parameter estimation objective function is in a quadratic form which means the confidence regions are exactly hyper-ellipsoids. For models that are non-linear in parameters, it is very difficult to determine exact parameter uncertainties since $R(\boldsymbol{\theta})$ is not a quadratic function regarding $\boldsymbol{\theta}$. Usually it will be approximated by the second order Taylor series expansion of $R(\boldsymbol{\theta})$ around the estimated parameters $\hat{\boldsymbol{\theta}}$, given in (2.6).

$$R(\boldsymbol{\theta}) \approx R(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \times \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \times \frac{\partial^2 R}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \times \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) \tag{2.6}$$

It should be noted that the best-fit parameters must be very close to the true parameter values so that this approximation is applicable. The first order derivative term is ignored in (2.6) because its value is zero at the minimum. Combining (2.5) and (2.6), the approximated confidence region can be obtained as:

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \times \mathbf{V}^{-1}(\hat{\boldsymbol{\theta}}) \times (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq p \times F_{p,N-p}^{1-\alpha} \tag{2.7}$$

where

$$\mathbf{V} = 2 \times \frac{R(\hat{\boldsymbol{\theta}})}{N-p} \times \mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}),\ \mathbf{H}\left(\hat{\boldsymbol{\theta}}\right) = \frac{\partial^2 R}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}^T} \tag{2.8}$$

Here $\mathbf{V}$ is the parameter estimation error covariance matrix which is the cornerstone to measure parameter estimation uncertainty. The diagonal elements of the matrix are the variances of parameter estimates and the off-diagonal elements are the covariance between parameters. $R(\hat{\boldsymbol{\theta}})/(N-p)$ is an approximation of residual variance σ^2 and $\mathbf{H}(\hat{\boldsymbol{\theta}})$ is the Hessian matrix which is determined by the second order derivative of $R(\hat{\boldsymbol{\theta}})$ with respect to parameters. Therefore, the confidence region of a single parameter θ_i, which is a one dimensional interval, can be determined by:

$$\delta_i = \pm t_{N-p}^{\alpha} \times \sqrt{\mathbf{V}_{ii}} \tag{2.9}$$

where t_{N-p}^{α} is the student distribution with $(1-\alpha)$ confidence level and $(N-p)$ degrees of freedom.

2.4 Parameter sensitivity analysis

Parameter sensitivity analysis studies how sensitive the model outputs are to changes in parameters. It can be used in the gradient based optimisation process for parameter estimation, or as the basis for most parameter subset selection and model simplification, or as an important component to construct FIM which can be further used in experimental design (Saltelli et al., 2008, 2004; Turányi, 1990). Therefore, parameter sensitivity plays an indispensable role in parameter estimation, parameter identifiability analysis and experimental design. Generally, parametric sensitivity analysis can be divided into two large categories: LSA and GSA. The former refers to small and individual changes in parameters, while the latter is focused on large magnitude changes of parameters and also considers parameter correlations. Both sensitivity analysis methods are useful for the investigation of the effects of parameters on model outputs. In this section, methods to calculate sensitivities are reviewed. Sensitivity analysis and its roles in model identification will also be discussed.

2.4.1 Local sensitivity analysis

Considering the general non-linear dynamic model in (2.1), denote x_i as the i-th state variable in $\mathbf{X}$ and θ_j the j-th parameter in $\boldsymbol{\theta}$, using Taylor series expansion, the change of x_i with respect to the change of θ_j can be expressed as:

$$x_i(\theta_j + \Delta\theta_j, t) = x_i(\theta_j, t) + \frac{\partial x_i}{\partial \theta_j}\Delta\theta_j + \frac{1}{2}\frac{\partial^2 x_i}{\partial \theta_j^2}(\Delta\theta_j)^2 + \cdots \tag{2.10}$$

where $\Delta\theta_j$ is a small variation for θ_j. The first order partial derivative $\partial x_i/\partial\theta_j$ is called parameter sensitivity. The analytic solution of the sensitivity can be obtained if the analytic solution of model equation (2.1) is known. This rarely happens for non-linear models and numerical methods have to be used to calculate the local sensitivity. A number of methods have been developed for the numerical calculation of local sensitivities, such as the finite difference method (FDM), the Green function method (GFM), the analytically integrated Magnus method (AIM), the direct differential method (DDM), the polynomial approximation method and others (Cukier et al., 1978; Dickinson and Gelinas, 1976; Yue et al., 2006; Zou and Ghosh, 2006). The most intuitive method is the FDM. For instance, the centre difference method is given as:

$$\frac{\partial x_i}{\partial \theta_j} \approx \frac{x_i(\theta_j + \Delta\theta_j, t) - x_i(\theta_j - \Delta\theta_j, t)}{2\Delta\theta_j} \tag{2.11}$$

where $\Delta\theta_j$ is an infinitesimal variation of the j-th parameter. Each time only one parameter is changed and the other parameters are kept at their nominal values. This is computationally tedious especially when there are a large number of parameters and state variables. Moreover, the variation of parameters needs to be chosen. Theoretically $\Delta\theta_j$ should be selected as small as possible, but in practical situations this will cause numerical inaccuracy, while large value of $\Delta\theta_j$ will cause the effect of model non-linearity with increased efforts and cost. The GFM and AIM methods use integration rather than differentiation to calculate sensitivities. They are more preferred when the number of parameters is much larger than the number of state variables.

The DDM, which is applicable to ODE models, has been developed to calculate the local sensitivities (Atherton et al., 1975). This method is stable and computationally efficient especially in computer simulation. The DDM will be used throughout this research work to calculate local parametric sensitivities. The derivative of parameter sensitivity with respect to time can be obtained from partial differentiation of ODEs in (2.1):

$$\frac{d}{dt}\frac{\partial \mathbf{X}}{\partial \boldsymbol{\theta}} = \frac{\partial \mathbf{f}}{\partial \mathbf{X}}\frac{\partial \mathbf{X}}{\partial \boldsymbol{\theta}} + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} = \mathbf{J}\cdot\mathbf{S}+\mathbf{F} \tag{2.12}$$

where $\mathbf{J} \in \mathbb{R}^{n\times n}$, $\mathbf{F} \in \mathbb{R}^{n\times p}$ and $\mathbf{S} \in \mathbb{R}^{n\times p}$ are Jacobian matrix, parametric Jacobian matrix and parameter sensitivity matrix which can be determined by:

$$\mathbf{J} = \frac{\partial \mathbf{f}}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}, \ \mathbf{F} = \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_1}{\partial \theta_2} & \cdots & \frac{\partial f_1}{\partial \theta_p} \\ \frac{\partial f_2}{\partial \theta_1} & \frac{\partial f_2}{\partial \theta_2} & \cdots & \frac{\partial f_2}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial \theta_1} & \frac{\partial f_n}{\partial \theta_2} & \cdots & \frac{\partial f_n}{\partial \theta_p} \end{bmatrix} \tag{2.13}$$

and

$$\mathbf{S} = \frac{\partial \mathbf{X}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{n,p} \end{bmatrix}, \ s_{i,j} = \frac{\partial x_i}{\partial \theta_j} \tag{2.14}$$

The n state equations in (2.1) are combined with $n \times p$ parameter sensitivity equations (2.12) to form into coupled differential equations:

$$\begin{aligned} \dot{\mathbf{X}}(t) &= \mathbf{f}(\mathbf{X}(t), \boldsymbol{\theta}),\ \mathbf{X}(t_0) = \mathbf{X}_0 \\ \dot{\mathbf{S}}(t) &= \mathbf{J} \cdot \mathbf{S}(t) + \mathbf{F},\ \mathbf{S}(t_0) = \mathbf{S}_0 \end{aligned} \tag{2.15}$$

where $\mathbf{S}_0$ is the initial condition of local parametric sensitivities which are normally zero unless they are connected to the initial condition of the state variables.

Both $\mathbf{X}(t)$ and $\mathbf{S}(t)$ can be determined simultaneously through direct integration of ODEs in (2.15). In most biochemical systems the divergence between parameter values is usually very large. The state variables may also have values at widely different scales (Cho et al., 2003; Yue et al., 2006; Zi, 2011). In order to compare the sensitivities from different state variables and across different parameters, the following relative (or scaled) sensitivity is normally used instead of the absolute sensitivities $s_{i,j}$.

$$\bar{s}_{i,j} = \frac{\partial x_i}{\partial \theta_j} \times \frac{\Delta \theta_j}{\Delta x_i} \tag{2.16}$$

There are several ways to define Δx_i and $\Delta \theta_j$, e.g. if preliminary measurement data is available, Δx_i can be chosen as the covariance of measurement error for the corresponding state variables, and $\Delta \theta_j$ can chosen as the uncertainty range of that parameter. If the covariance of measurement error and the uncertainty range of the parameter are not known, Δx_i can be chosen as the average value of each state variable across the whole experimental time and $\Delta \theta_j$ can be chosen as its nominal value.

Apart from (2.16), the logarithmic function is another way to eliminate the magnitude difference between model parameters or across model outputs, from which the relative parameter sensitivities can be determined as

$$\bar{s}_{i,j} = \frac{\partial x_i}{\partial \theta_j} \times \frac{\theta_j}{x_i} = \frac{\partial \ln x_i}{\partial \ln \theta_j} \tag{2.17}$$

In most biochemical systems the difference of values between model parameters can be up to several orders, the log function is quite useful in these cases. However, (2.17) is not applicable when x_i is numerically zero. The comparison of relative sensitivity calculations (2.16) and (2.17) will be discussed in Chapter 4.

The parametric sensitivities are time dependent. To calculate the overall effects along the whole time scale, the following measures can be used. Taking the relative sensitivities as an example, once the numerical solution of local sensitivity matrix is determined, the overall effect of parameter θ_j to state variable x_i can be calculated by

$$RS_{i,j} = \frac{1}{N}\sqrt{\sum_{l=1}^{N} \bar{s}_{i,j}(t_l)^2} \tag{2.18}$$

To look at the overall effect of the parameter to multiple states, the following measurement is taken:

$$OS_j = \frac{1}{N}\sqrt{\sum_{i=1}^{n}\sum_{l=1}^{N} \bar{s}_{i,j}(t_l)^2} \tag{2.19}$$

Following results by LSA, model parameters can be grouped into important parameters and less important parameters. For those less important parameters, due to their limited effects to model outputs, they can be fixed at constant values or even removed from the model in order to simplify the model (Chu and Hahn, 2007). It is expected that removing these parameters does not affect the remaining parameter sensitivities (Brun et al., 2001). However, in practice, the important parameters could be affected as they are often correlated to those removed parameters. This requires parameter identifiability analysis that considers both parameter effects and correlations between them. Nevertheless, the LSA provides an intuitive understanding of parameter effects to the systems and it is closely related to most parameter subset selection approaches. In addition, when the gradient-based or Hessian-based numerical algorithms are used to solve a parameter estimation problem, the local parameter sensitivity is essential for the optimisation routine. The broadly used model-based OED requires the calculation of the information content which depends on the calculation of local sensitivity coefficients.

2.4.2 Global sensitivity analysis

When the parameter values are only approximately known, which is often the case in biochemical systems, large parameter uncertainties need to be considered in sensitivity analysis. GSA can be used to investigate the non-linear effects and correlations between parameters over a large uncertainty region around nominal parameter values (Chu et al., 2010; Jin et al., 2007).

Different from LSA in which only one parameter is changed around its nominal value, GSA incorporates the effect of parameters that are varying simultaneously. This helps in finding the degree of parameter interactions and facilitates model building by determining important parameters which contribute most to the model output uncertainties. Several methods have been developed to calculate GSA, such as the FAST method and the Sobol's method (McRae et al., 1982; Saltelli and Bolado, 1998; Turányi, 1990). Detailed reviews of GSA can be found in literature (Iooss and Lemaître, 2015; Kucherenko, 2005; Saltelli, 2004; Saltelli et al., 2008; Sobol, 2001). In this work, the Morris screening method and the Sobol's method are employed. GSA calculation in biochemical system models and its connection to the OED will be discussed in appendix B.

2.5 Identifiability analysis

The number of model parameters depends on the scale and the complexity of systems. In most cases, not all the model parameters can be reliably estimated through system identification due to the lack of parameter identifiability. Parameter identifiability analysis investigates whether model parameters can be estimated from experimental data. In this section, the identifiability analysis is reviewed on both the structural identifiability and the practical identifiability. The structural identifiability is introduced first. Then methods for the analysis of practical identifiability and their recent development are presented. The parameter subset selection methods will be discussed next based on identifiability analysis.

2.5.1 Structural identifiability

One major issue in parameter estimation of complex models is that the parameters are not identifiable. It is necessary to determine whether this is caused by inappropriate model structure or insufficient and imprecise experimental data. The structural identifiability analysis can uncover problems with model structure and figure out whether it is possible to obtain unique parameter values from perfect noise-free data (Chis et al., 2011; Walter and Pronzato, 1996; Zhang et al., 2010). If the parameters can be uniquely estimated from noise-free experimental data, then the model and parameters are said to be structurally identifiable. Consider the general dynamic model in (2.1), if model parameter vector $\boldsymbol{\theta}$, meet the condition that (Ljung and Glad, 1994)

$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p, \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_1) = \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_2) \Leftrightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \tag{2.20}$$

Then the parameters $\boldsymbol{\theta}$ are said to be globally identifiable. If the condition holds only for a neighbourhood of the nominal parameters $\boldsymbol{\theta}^*$ which is given by

$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \{\boldsymbol{\theta} \in \mathbb{R}^p | \ \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \boldsymbol{\delta}\}, \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_1) = \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_2) \Leftrightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \qquad (2.21)$$

Then the parameters $\boldsymbol{\theta}$ are said to be locally identifiable. It is straightforward to determine the identifiability of models that are linear in parameters.The identifiability of non-linear models is much more difficult to determine. Several methods have been developed for the structural identifiability of non-linear models. A method has been proposed in which only the linearised version of the non-linear model is analysed to check the identifiability (Grewal and Glover, 1976). If the linear part is identifiable, then the complete non-linear model is also regarded structurally identifiable. This method is simple and efficient for some models. However, when the linearised part of the model is non-identifiable, it gives no clue about whether the full non-linear model is identifiable (Ben-Zvi et al., 2006). Other techniques such as the Taylor series expansion approach (Grewel and Glover, 1976; Pohjanpalo, 1978), the generating series method (Walter and Lecourtier, 1982), the similarity transformation approach (Vajda et al., 1989; Vajda and Rabitz, 1989), and the differential algebra algorithm (Audoly et al., 2001), are also developed for structural identifiability analysis. Most of the proposed methods can be applied mainly to systems with moderate complexity.

2.5.2 Practical identifiability

When model parameters are deemed as structurally identifiable, they may not be practically identifiable due to several reasons: (1) experimental data for parameter estimation are sparse and noisy, or contains limited information due to poorly designed experiments; (2) some parameters have very little influences on model outputs; (3) the effect of some parameters on the model prediction can be compensated by other parameters (Brockmann et al., 2008;

Huang et al., 2010; Wu et al., 2008), in another words, high correlations exist between parameters. Therefore, both sensitivity of model predictions to changes in parameter values and correlations between parameters need to be considered to assess practical identifiability.

For complex biochemical process models with high non-linearity and a considerable number of parameters, it is often difficult to identify all parameters from available data. It is critical to find the identifiable parameter subset given a measured data set. The model can be simplified by eliminating some non-identifiable parameters, or by lumping several parameters together, or by setting non-identifiable parameters at fixed values. During the parameter estimation procedure, the focus is then put on those identifiable parameters that are important to model predictions. Numerically, the parameter subset selection problem is a discrete (combinatorial) non-convex optimisation problem. Exhaustive search and genetic algorithm (GA) are the most widely used methods to solve such problems. However, for non-linear dynamic systems with a large number of parameters, these methods are computationally expensive. Therefore, several approaches have been developed to evaluate practical identifiability. Five useful methods are briefly presented in the following.

(i) *Visual inspection of local sensitivities*

By checking local sensitivity functions, the influences of parameters on outputs and also the correlations between parameters can be visually determined. This knowledge can be used to select the identifiable parameter subset (Holmberg, 1982; Petersen et al., 2001; Reichert and Vanrolleghem, 2001; Seagren et al., 2003). Note that appropriate scaling is important when performing LSA to make sure that the sensitivities are compared at the same scale. This method is intuitive, however, it can be intractable for large complex models because higher order interrelationships among three or more parameters cannot be visually detected.

(ii) *Collinearity index method*

Another technique was developed which is based on the scaled sensitivity matrix and correlations between the effects of two or more parameters (Brun et al., 2001). A re-scaled sensitivity measure is

$$\tilde{s}_{i,j} = \frac{\bar{s}_{i,j}}{\sqrt{\sum_{l=1}^{N} \bar{s}_{l,j}^2}} \tag{2.22}$$

A measure for the importance of a single parameter to all the state variables can be described as:

$$\delta_j^{msqr} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \tilde{s}_{i,j}^2} \tag{2.23}$$

Parameters with higher δ^{msqr} values have more significant effects on model outputs. In the next step, all possible parameter combinations will be considered. Taking $\tilde{s}_{i,j}$ to form the sensitivity matrix, $\tilde{\mathbf{S}}$, a collinearity index for each parameters subset will be calculated in order to determine the linear dependencies of their sensitivity functions:

$$\gamma(\boldsymbol{\theta}) = \frac{1}{\lambda_{min}\left(\tilde{\mathbf{S}}(\boldsymbol{\theta})^T \tilde{\mathbf{S}}(\boldsymbol{\theta})\right)} \tag{2.24}$$

where λ_{min} is the smallest eigenvalue of a matrix. The values of $\gamma(\boldsymbol{\theta})$ range from unity, for linearly independent parameter sensitivities, to infinity, for increasing degrees of dependency between parameter sensitivities. A threshold of 10-15 for the value of $\gamma(\boldsymbol{\theta})$ is suggested to determine whether a parameter subset is identifiable or not (Brun et al., 2001). However, this method could be computationally expensive for models with a large number of parameters. Also there are normally more than one parameter subset that fulfill the threshold of $\gamma(\boldsymbol{\theta})$ and it is unclear which parameter subset should be selected for parameter estimation.

(iii) *Methods based on principal component analysis and eigenvalues*

Two parameter subset selection techniques have been proposed based on principal component analysis (PCA) (Degenring et al., 2004; Li et al., 2004). The scaled parameter sensitivities are first calculated based on (2.16), and eigenvalues and eigenvectors can be obtained for matrix $\bar{\mathbf{S}}^T\bar{\mathbf{S}}$. The absolute value of the principal component element reflects the contribution of the j-th parameter to the variance of the i-th output. The proposed index contains information about parameter importance with respect to model outputs and the linear dependency of the parameter on other parameters. A different way to analyse parameter identifiability has been proposed in (Degenring et al., 2004). Three different sequential procedures are suggested to select identifiable parameters by comparing eigenvalues and their corresponding eigenvectors. Parameters are ranked either from least estimable to most estimable (backward selection), or from most estimable to least estimable (forward selection), based on their corresponding weights in the eigenvector. This method is easy to implement but different results may be obtained by using different procedures, and it also requires an arbitrary threshold value.

(iv) *Methods based on scalar measure of FIM*

A new parameter subset selection technique has been developed based on the scalar measure of FIM because the FIM contains the sensitivity function and measurement uncertainties (Weijers and Vanrolleghem, 1997). A three step procedure is proposed to determine identifiable parameters. In the first step, the scaled parameter sensitivities are calculated and the overall effects of parameters on model outputs are compared based on (2.16) and (2.19). Parameters with small influence on model predictions will be ignored and the rest are maintained in the model. In the second step, the FIMs were constructed for all possible parameter subsets of the surviving parameters and the D-optimality and the modified E-optimality are calculated and compared. The final step is to select the number of parameters to be estimated based on the two scalar measures of FIM calculated in the second step. There is no criterion to select an appropriate number in the parameter estimation subset. Instead, the authors used their expert experience to select as many parameters as possible so long as it will not cause numerical problems during parameter estimation. This method could also be computationally expensive in the second step for high dimensional models.

(v) *Orthogonalisation based methods*

The orthogonalisation method for selecting estimable parameters was originally proposed in a complex dynamic model for *ethylene-butene* co-polymerisation with 50 parameters (Yao et al., 2003). This algorithm takes account of the net influence of model parameters on response variables and correlations with other parameters. The estimable parameters are selected through a sequential process. This method has been used in several chemical and biological models.

One problem with this orthogonalisation based method is that by using the forward selection algorithm (or backward elimination), a non-optimal solution may be obtained. During the parameter subset selection procedure, the important parameters are selected step by step. At each step the parameter ranking and selection is based on the previous selected parameters. In this way the final identified parameter set might not be the best combination for parameter estimation. Recently, a new method based on parameter clustering was proposed which allows selection of multiple parameters in each iterative step and this method is particularly helpful to deal with large biochemical systems with hundreds to thousands of parameters (Chu and Hahn, 2008; Chu et al., 2009). An efficient algorithm based on binary search was also developed which makes the set by set selection possible (Alberton et al., 2013). It was demonstrated that based on the sensitivity matrix, each time the whole parameter set can be divided in two subsets, and identifiability of each subset can then be tested. If one subset is identifiable, then one can add half of the parameters from another subset into this subset to further test the identifiability. This process can be performed iteratively until the largest identifiable parameter subset is obtained. Another problem associated with the orthogonalisation method is that the local sensitivities are based on the initial parameter values. Bad choice of initial parameters may cause erroneous conclusions for parameter ranking. In order to solve this problem, different initial parameter values within the uncertainty ranges of parameters are suggested for the parameter ranking and the best set that most regularly appeared in each ranking process is chosen as the suitable parameter set (McLean et al., 2012).

Another criterion based on mean square error has been proposed to determine the number of parameters for estimation instead of setting a cut-off value (Wu et al., 2011a,b). The balance between model prediction and precision of parameter estimation is considered. The number of estimated parameters is determined when the increase of parameter estimation bias is equal to or lower than the decrease of parameter estimation residual. However, this method requires real experimental information. A Monto Carlo method combined with the orthogonalisation based method has been developed which is used to evaluate the average bias by using a lot of different initial parameter values and to find the best model prediction (Chu et al., 2009). This method is stated to be more robust to poor initial parameter guesses. Both methods involves parameter estimation which requires a considerable computational effort. Most recently, the mean square error based criterion has been improved in order to make an accurate model prediction by considering different experimental conditions (Eghtesadi and McAuley, 2014, 2016).

It can been seen that all the above parameter subset selection methods are based on the averaged effect of model parameters on model outputs, and the correlations between parameter pairs. The parameter subset selection approaches will be investigated in Chapter 4. Their importance in model-based OED will be justified through the implementation to two biochemical exemplar systems.

2.6 Summary

In this chapter, the model building and validation procedure for biochemical systems has been reviewed. The LSE algorithm and challenges in parameter estimation are presented. The (parametric) LSA is reviewed as it is important during the whole model building procedure. The GSA is also briefly introduced as its potential use in the model-based OED will be discussed for systems with model uncertainties (see Appendix C). Finally, the parameter identifiability is studied. Both the structural identifiability analysis and the practical identifiability analysis are reviewed. The parameter subset selection approaches have been discussed. Overall, this chapter provides a review of the preliminary knowledge necessary for model identification and OED. In this PhD work, the DDM algorithm will be used to calculate parameter sensitivities. The generating series method will be applied to examine the structural identifiability of case study models. The orthogonalisation based methods and the PCA based method will be used to figure out important and identifiable model parameters.

Chapter 3

Fundamentals and Applications in Optimal Experimental Design

Experimental design plays important roles in model building and validation process, as discussed in Chapter 2. The purpose of OED is to design experiments in such a way that the experimental data contains as much information as possible. In the OED process, one needs to decide when, where and how the system should be driven or manipulated, and when, where and how the measurement should be taken during experimentation.

From the modelling perspective, the priority is to obtain a precise model which is able to describe the real process under various conditions. Therefore, the major interest of OED is on system identification, including model discrimination and parameter estimation. The design of experiments should help to achieve better understanding of the system under study, or to increase the ability of model prediction at specific experimental conditions. Ideally model identification and process control and optimisation should be handled together. However, the statistical experimental designs are not able to meet this requirement due to the increased scale of systems and their complex dynamics (Dejaegher and Vander Heyden, 2011; Georgakis, 2013; Hibbert, 2012). It is necessary to develop new OED techniques which can provide more insightful information for modelling, analysis and operation of real processes. This motivates the work on model-based OED in this thesis.

In this chapter, the experimental design techniques and recent applications are reviewed. Firstly, a brief review of the statistical experimental design is given in Section 3.2 and the main drawbacks of this technique are discussed through the comparison with model-based OED. In Section 3.3, the model-based OED for both model discrimination and parameter estimation are discussed. A review of OED for model discrimination and recent developments are presented. Then the OED for parameter estimation is described in detail. Basic concepts, such as OED formulation, design factors, and scalar design criteria are studied which support the further OED development in the thesis. The development of OED by considering different design factors is also provided. In Section 3.4, the numerical aspect of solving OED problems is discussed. A summary is given in Section 3.5.

3.1 Introduction

The basic idea of experimental design can be dated back to 1936 when Fisher described the experimental design problem as deciding the experimental factors most influential on variables and optimising combinations of experimental factors so that the outputs of interest are optimised (Fisher, 1936). This early statistical experimental design is mainly concerned with the relationship between the input factors and the output responses for linear, static systems. The design process is to evaluate several experimental factors at selected levels in order to decide the best experimental set-up. Statistical design techniques are straightforward to implement and the results are easy to interpret. They have been used in biochemical and biological area (Fiordalis and Georgakis, 2013; Fricke et al., 2013; van Riel, 2006b).

Another type of experimental design is called model-based OED, which has received more and more attention in recent studies, especially in biological and biochemical fields because of its ability to promote fast development, refinement and statistical validation of linear or non-linear, static or dynamical models. The key advantages of model-based OED over statistical design are summarised as follows: (1) the explicit use of prior model information, e.g. model structure, initial model parameter information, and priori data information; (2) applicable for complex systems design, e.g. systems with non-linear, dynamic, high dimensional characteristics; (3) OED investigation with different or multiple objectives through different criteria; (4) numerical solution can be obtained through an optimisation framework. In the following sections, the statistical design and the model-based OED will be reviewed, separately.

3.2 Brief review of statistical experimental design

Statistical design of experiments refers to the process of planning the experiment in such a way that appropriate data can be collected which can be analysed by the statistical methods to achieve a proper input-output. To use statistical approaches in the design and analysis of an experiment, the overall objective needs to be defined. The specific questions to be addressed during experimentation are related directly to the defined overall objective. A key step of statistical design is to select the responses and the input factors at pre-specified levels. The selected responses need to include useful information about the process under study. The potential input design factors that can be varied in the experiment should also be specified from the beginning. It is important to investigate key influential factors and this is where process knowledge is required. Once the design factors are determined, the ranges or the levels of these factors will be designed.

There are several statistical design methods developed. One classically applied strategy is the one-factor-at-a-time (OFAT) approach, in which only one factor is varied and optimised each time while keeping the other factors constant. This method is simple but requires a lot of experimental work when there is a considerable number of factors. Also, it fails to consider any possible correlations between the factors. An alternative to the OFAT method is the so-called factorial design which is one of the main approaches for the statistical design of experiments (Dejaegher and Vander Heyden, 2011). A factorial design involves two or more factors, each with discrete possible values or levels. The experimental strategies take on all possible combinations of these levels across all the factors. In this full factorial design, the number of runs required increases exponentially as the number of factors of interest increases. A modified fractional factorial design was proposed which simplifies the basic full factorial design. Only a subset of factors are investigated which allows detection of interactions among design factors (Hibbert, 2012).

The statistical design can be roughly classified into screening design and response surface design. The screening design is used to identify the most influential factors by screening a relatively large number of factors in a small number of experiments. The response surface design is to find the optimal levels of the most important factors by examining experimental factors at different levels. All these methods are only focused on the influences of experimental factors to the output responses while the underlying model information are not considered. Also, in the statistical experimental designs the prior knowledge is not exploited which may cost experimental efforts. It is therefore essential to develop more advanced experimental design methods which can effectively cope with complex dynamic systems; the so-called model-based experimental design techniques are thus developed.

3.3 Model-based experimental design

The model-based OED is focused on acquiring the most informative measurement data through the designed experimental strategies with reduced experiment efforts. It allows more design freedoms than does statistical experimental design. The statistical experimental designs, treating the studied model as a ‘black box’, in most cases only focus on the design of influential input factors, while the model-based OED is able to incorporate both input design factors and measurement strategies (such as sampling time schedule, measurement set and experimental length) into an integrated design framework over a wide range of experimental conditions. Therefore, the model-based OED techniques have attracted increased attention in science and engineering, particularly in chemical systems and systems biology. In this section, the model-based OED approaches for parameter estimation and model discrimination are reviewed which provides necessary theoretical support for follow-on OED development. The single factor OED for the input intensity, the sampling time profile and the measurement set are also described.

3.3.1 Experimental design for model discrimination

Model discrimination is the procedure to identify the most appropriate model from a set of rival models. The residuals from parameter estimation are contributed by measurement error and lack of fit of the model. The OED for model discrimination aims at maximising the divergence of model prediction between rival models attributed to lack-of-fit, so that the most suitable model can be selected (Wiens, 2009). As depicted in Figure 3.1, the OED for model discrimination contains several steps. Firstly, a model fitting process needs to be conducted for all model candidates by using available experimental data. Then the model's adequacy to describe the available data is tested in order to find out which model(s) is able to describe the process with a reasonable quality. Next the optimisation design based on certain criteria is implemented for those models that pass the adequacy test in the previous step. This optimised experiment is then performed and new data will be generated, with which parameter re-estimation and re-evaluation can be made for all adequate models. The process is iterated until the best model is identified. It should be noted that if all candidate models are identified to be inadequate, new models need to be proposed. Also, it is possible that more than one model is found to be suitable to describe a system and it is difficult to discriminate among them through OED. In such a case, the best model is selected as a trade-off between model fit, model simplicity and identifiability. OED for model discrimination can be formulated as an optimisation problem where the objective function, denoted as *D*, is

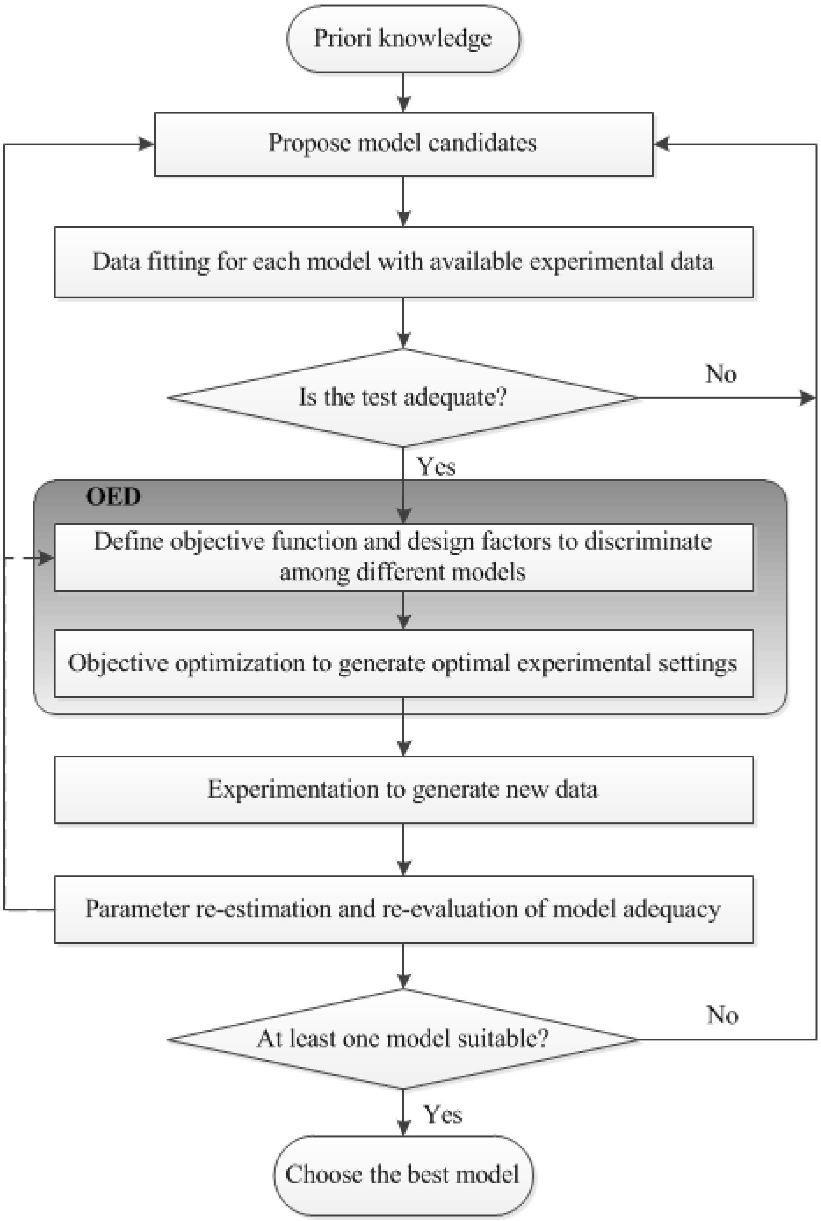

Priori knowledge
Propose model candidates
Data fitting for each model with available experimental data
Is the test adequate?
No
Yes
OED
Define objective function and design factors to discriminate among different models
Objective optimization to generate optimal experimental settings
Experimentation to generate new data
Parameter re-estimation and re-evaluation of model adequacy
At least one model suitable?
No
Yes
Choose the best model

Fig. 3.1 Model discrimination procedure based on OED

the divergence of output predictions between different models.

$$\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} D(\boldsymbol{\phi}) \tag{3.1}$$

where $\boldsymbol{\phi}$ are the experimental degrees of freedom which are restricted by constraints that define the admissible set of experiments. This objective function forms the basis of discriminatory experiments and there are several design criteria based on (3.1) which have been proposed and successfully applied in chemical engineering and other systems. Here a brief review is given to some most widely used design criteria for model discrimination.

The measurement data from experimental design should be as different as possible between different models to allow for model discrimination. An objective function is formulated in (Hunter and Reiner, 1965), in which the expected experiment set $\boldsymbol{\phi}$ should be selected so as to maximise the divergence of predictions between two models.

$$\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \left(\hat{\mathbf{Y}}_1(\boldsymbol{\phi},\boldsymbol{\theta}_1) - \hat{\mathbf{Y}}_2(\boldsymbol{\phi},\boldsymbol{\theta}_2)\right)^2 \tag{3.2}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two different parameter sets of models 1 and 2, respectively; $\hat{\mathbf{Y}}_1(\boldsymbol{\phi},\boldsymbol{\theta}_1)$ and $\hat{\mathbf{Y}}_2(\boldsymbol{\phi},\boldsymbol{\theta}_2)$ are the predicted responses of models 1 and 2 along the whole sampling time points and across all measured output variables, respectively. It is assumed that the variances of model outputs are the same throughout the whole experiment region. It is impossible to distinguish model predictions when model output uncertainties are at different levels, even with a high divergence of the model responses. The design criterion can be easily modified by considering measurement errors and the improved design problem is formulated as:

$$\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}\in\Phi} \left(\hat{\mathbf{Y}}_1(\boldsymbol{\phi},\boldsymbol{\theta}_1) - \hat{\mathbf{Y}}_2(\boldsymbol{\phi},\boldsymbol{\theta}_2)\right)^T \mathbf{Q}^{-1} \left(\hat{\mathbf{Y}}_1(\boldsymbol{\phi},\boldsymbol{\theta}_1) - \hat{\mathbf{Y}}_2(\boldsymbol{\phi},\boldsymbol{\theta}_2)\right) \tag{3.3}$$

where $\mathbf{Q}$ is the measurement error covariance matrix. The inverse of $\mathbf{Q}$ is used as the weighting matrix.

A design criterion for model discrimination which is an extension of Hunter and Reiner's work (Hunter and Reiner, 1965) was developed (Buzzi-Ferraris and Forzatti, 1983). Both the measurement errors and the uncertainty of output predictions are considered where the weighting matrix in (3.3) is modified to represent the uncertainty in the difference between the predicted outcomes of the two rival models. Later on, the criterion is extended into multi-output models and the maximum discriminant value must be larger than the number of output variables; otherwise the model discrimination is not possible due to the χ^2-test for the model adequacy (Ferraris et al., 1984).

A design criterion, called T-optimality, has been proposed which is based on the D-optimality design theory used for discrimination between two rival regression models (Atkinson, 2008; Atkinson and Fedorov, 1975a,b). Suppose that the functions of the two rival models are $\mathbf{h}_1\left(\mathbf{X}(t),\boldsymbol{\theta}_1\right)$ and $\mathbf{h}_2\left(\mathbf{X}(t),\boldsymbol{\theta}_2\right)$, respectively, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the unknown parameter vectors. Assuming the first model is the true model, the observations are given by

$$\mathbf{Y}=\mathbf{h}\left(\mathbf{X}(t),\boldsymbol{\theta}\right)+\boldsymbol{\xi}(t)=\mathbf{h}_1\left(\mathbf{X}(t),\boldsymbol{\theta}_1\right)+\boldsymbol{\xi}(t) \tag{3.4}$$

The measurement error $\boldsymbol{\xi}$ is assumed to be independently, normally distributed with zero mean and constant variance σ^2 for each channel. Then the experiment is designed to find the maximum value of the sum of squares for the lack of fit of the second model $\mathbf{h}_2$, which is given by

$$\begin{aligned}\Delta_1(\boldsymbol{\phi}) &= \sum_{i=1}^{n}\omega_i\left\{h(x_i)-h_2(x_i,\hat{\boldsymbol{\theta}}_2)\right\}^2 \qquad (3.5)\\ \boldsymbol{\phi} &= \left\{\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ \omega_1 & \omega_2 & \cdots & \omega_n \end{array}\right\}\end{aligned}$$

where

$$\sum_{i=1}^{n} \omega_i \left\{ h(x_i) - h_2(x_i, \hat{\boldsymbol{\theta}}_2) \right\}^2 = \inf_{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2} \sum_{i=1}^{n} \omega_i \{ h(x_i) - h_2(x_i, \boldsymbol{\theta}_2) \}^2 \tag{3.6}$$

The design factors, $\boldsymbol{\phi}$, which is called T-optimality criterion is aimed at maximising the difference of the second model from the first model which is assumed to be the ‘true’ model. The T-optimality can also be extended to the model discrimination between three or more models. With three or more proposed models, one design that can effectively detect the departure of one model may not be suitable for other models, so a weighting is defined to represent the importance of various departures. The T-optimal design is only suitable for models with homoscedastic error distribution and the ‘true’ model needs to be selected, while in practice it is not known which model is more adequate to be the true model. Although this design method has been extended into heteroscedastic error cases, the assumption of normal distribution noise is still needed (Ucinski and Bogacka, 2004).

An alternative sequential procedure for model discrimination was proposed which considers the model output uncertainties (Box and Hill, 1967). The authors introduced the entropy term as a measure of information for non-Gaussian systems and defined a discrimination function to maximise the change of entropy. The least possible information corresponds to the maximum entropy when all models have the same probability to be the true one. The main drawbacks of this design method are: (1) the updated probability of the model is dependent on the order of the experiments not on the available experimental information; (2) it is forced to make a selection among rival models even when all the candidates may be inadequate.

An alternative criterion called Kullback-Leibler (KL) optimality is proposed for model discrimination with non-Gaussian observations (López-Fidalgo et al., 2007). With this method, it is assumed that the probability density functions of the two rival models are known, based on which the KL distance can be calculated as a measure for model discrimination. The KL-optimal criterion is extended to a semi-parametric setup in which only the moment conditions are needed for the null or alternative regression models (Otsu, 2008). The Bayesian KL-optimal criterion is introduced in (Tommasi and López-Fidalgo, 2010) for statistical models with prior statistical information.

These design methods mentioned above are aimed at selecting the model which can best fit the experimental data of the system. It is worth to mention that many methods discussed in OED for parameter estimation are also applicable for model discrimination. Although different objective functions are used for these two different purposes, the numerical solution strategies applied and the statistical analysis tools adopted for model adequacy test and parameter significance are similar in both cases.

3.3.2 Experimental design for parameter estimation

When the most suitable model structure is determined from model discrimination, OED can be undertaken for parameter estimation. This is the main objective of this PhD work. The purpose of OED for parameter estimation is to devise efficient and necessary experimentation in order to generate the most informative measurement data that could reduce parameter estimation uncertainties. Several important aspects need to be considered such as the representation of data information, the experimental factors, and the effect of model parameters on output variables of interest. The general steps of OED for parameter estimation are given in Figure 3.2. To start with, the model prediction is fitted to the available experimental data to obtain rough estimates of model parameters. In the next step, the parameter identifiability is analysed to find the identifiable parameter subset. This parameter subset selection is essential because quite often it is not possible to estimate all model parameters for a complex system. OED involving a large number of parameters might result in non-informative experimental data. In the step on OED, the design criterion and the experimental factors are chosen based on expert experience and design purpose. A scalar function of FIM is thus constructed as a measure of data information. The numerical solution of the OED problem will provide optimal settings for the intended experimentation. The experiment is then performed using the optimised settings. New data is created and used for re-estimation of model parameters. The model with updated parameter values will be tested to check how well it can represent the real dynamic process. This parameter identification process is run iteratively until the satisfactory model parameters are obtained.

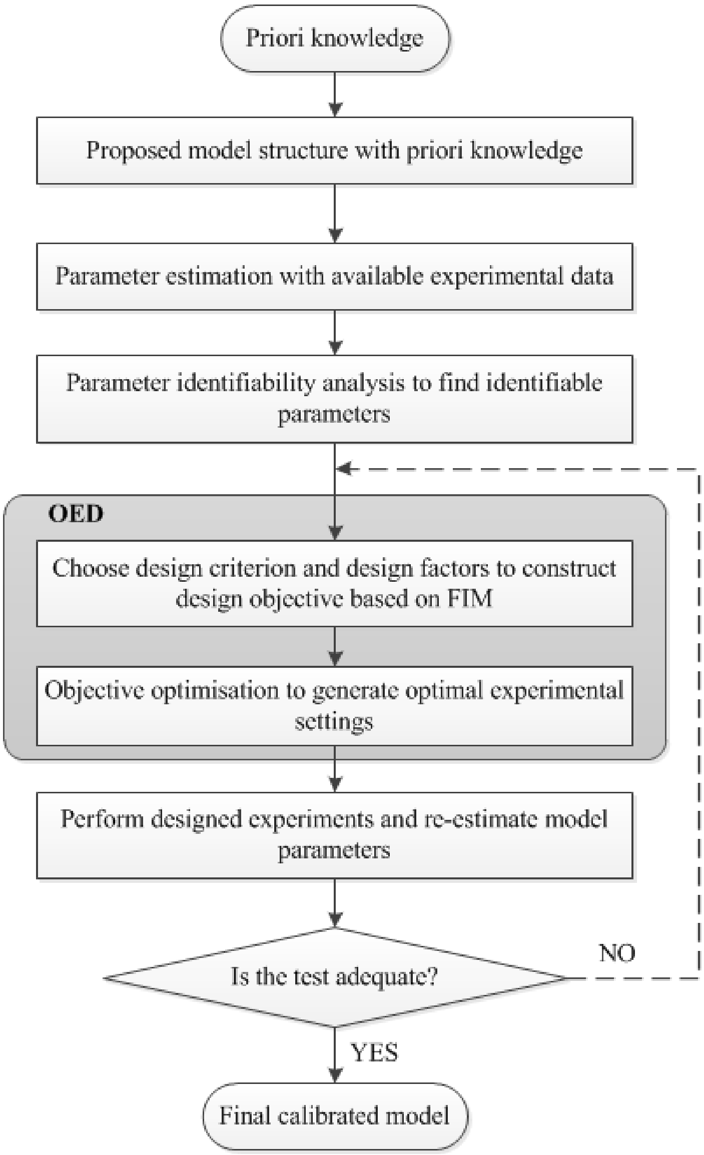

Priori knowledge
Proposed model structure with priori knowledge
Parameter estimation with available experimental data
Parameter identifiability analysis to find identifiable parameters
OED
Choose design criterion and design factors to construct design objective based on FIM
Objective optimisation to generate optimal experimental settings
Perform designed experiments and re-estimate model parameters
Is the test adequate?
NO
YES
Final calibrated model

Fig. 3.2 Parameter identification process based on OED

A comprehensive introduction on OED objective based on FIM, the scalar design criteria and several other related techniques are discussed in the following.

Objective function and design criteria

In model-based OED for parameter estimation the typical design factors can be collected into a design vector containing input and observation, given as:

$$\boldsymbol{\phi} = (\varsigma, \mathbf{T}_{sp}, \mathbf{z}) \tag{3.7}$$

where ς represents input factors, $\mathbf{T}_{sp}$ denotes the vector of sampling times and mostly the last sampling time shows the experiment duration, $\mathbf{z}$ represents the vector of selected measurement set. The constraints for all the design factors constitute the whole experimental design space $\boldsymbol{\Phi}$. Therefore, the OED for parameter estimation can be formulated as:

$$\arg\min_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \upsilon(\boldsymbol{\phi}, \boldsymbol{\theta}) \tag{3.8}$$

subject to model equation (2.1). Here $\upsilon(\cdot)$ is an objective function that represents the goodness of the experiments for parameter estimation. In the LSE framework, the parameter covariance matrix can be used to reflect parameter estimation quality. The FIM is widely used to approximate the inverse of the parameter estimation error covariance, which is defined as (Zullo, 1991):

$$\mathbf{FIM}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbf{S}(\boldsymbol{\theta}, \boldsymbol{\phi})^T \mathbf{W}\mathbf{S}(\boldsymbol{\theta}, \boldsymbol{\phi}) \tag{3.9}$$

where the weighting matrix $\mathbf{W}$ is normally taken to be $\mathbf{Q}^{-1}$ for most general discussion. The FIM can be used to quantify the information content of an experiment towards parameters to be estimated. The more sensitive of a state variable to a parameter, the more information is contained in the FIM about that parameter. The inverse of the measurement error covariance matrix, $\mathbf{Q}$, in the FIM indicates that a larger measurement error will contribute less reliable information than the data with a smaller measurement error. In addition, the correlations between measurements are also considered in the FIM. When the model is linear in its parameters, according to the Cramer-Rao lower bound inequality, the FIM is approximately equal to the inverse of the parameter estimation error covariance matrix, $\mathbf{\Sigma}$, under the assumption of unbiased parameter estimation and uncorrelated additive white measurement noise space(Ljung, 1998). Therefore, the local lower bound of the variance for the i-th parameter can be determined by (Schittkowski, 2007)

$$\delta_i^2 \geq \mathbf{\Sigma}_{ii} \tag{3.10}$$

The effect of OED on parameter estimation for a single parameter is illustrated in Figure 3.3. Two sets of experimental data are used for parameter estimation: the data set generated from the designed experiment (OED shown in blue curve) and the non-designed experimental data (red curve). It can be seen that the parameter estimation using the optimal experimental data can lead to a much narrower confidence interval than using the non-designed experimental data.This suggests that with OED, the generated experimental data can provide higher parameter estimation quality than non-designed condition.

In model-based OED, the objective function is normally expressed as some scalar quantities of **FIM** (or $\mathbf{\Sigma}$). The commonly used ‘alphabet’ optimisation criteria are introduced as follows (Balsa-Canto et al., 2007; Faller et al., 2003; Ljung, 1998)

- A-optimal: $\min\left\{trace\left(FIM^{-1}\right)\right\}$

  $trace(\cdot)$ is the summation of the diagonal elements of a matrix. The A-optimal design is to minimise the sum of eigenvalues of the estimation error covariance matrix. It can be taken as minimising the arithmetic mean of the parameter identification errors.

- D-optimal: $\max\left\{\det\left(FIM\right)\right\}$ or $\min\left\{\det\left(FIM^{-1}\right)\right\}$

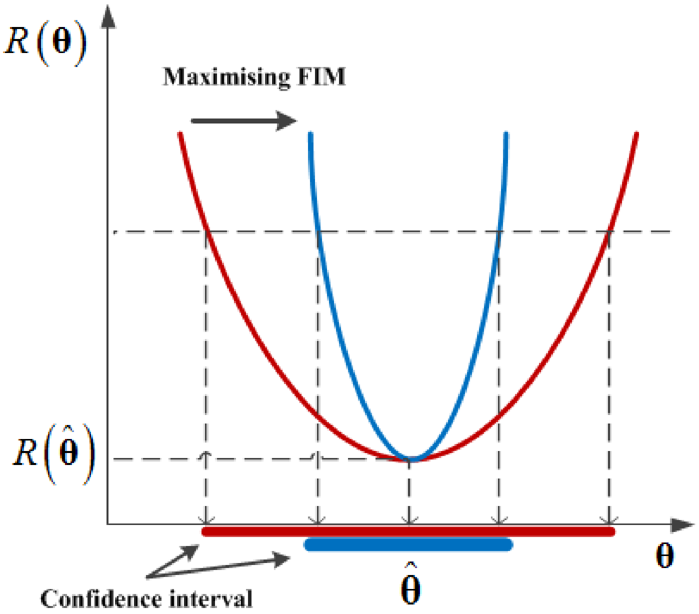


Fig. 3.3 Influence of FIM maximisation on the parameter estimation cost function and confidence intervals for a single parameter estimation problem

det$(\cdot)$ denotes the determinant function of a matrix. The D-optimal design is aimed at maximising the determinant of FIM in order to minimise the volume of the confidence ellipsoid for the estimated parameter errors in linear approximation.

- E-optimal: $\max\{\lambda_{min}(FIM)\}$ or $\min\{\lambda_{max}(FIM^{-1})\}$

  $\lambda_{min}(\cdot)$ is the minimum eigenvalue of a matrix. The E-optimality criterion secures the minimisation of the largest error of the covariance which is equivalent to minimise the longest axis of the confidence ellipsoidal.

- Modified E-optimal: $\min(\lambda_{max}(FIM)/\lambda_{min}(FIM))$

  $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ are the maximum and the minimum eigenvalues of a matrix. The modified E-optimal design aims at optimising the functional shape of the CIs as spherical as possible by minimising the ratio of the maximum to the minimum eigenvalues. In other words, the purpose of this design is to minimise the condition number of the information matrix.

As the FIM is approximately the inverse of the parameter estimation error covariance matrix, the above scalar measures are closely related to the shape, size and orientation of parameter estimation CI, as shown in Figure 3.4 for a two-parameter example. The yellow-shadowed area represents the confidence region of the parameters. It should be noted that these alphabetic criteria have some limitations and may not be directly applicable to specific systems. For example, the A-optimal criterion may result in a less informative experiment compared to other design criteria when the correlations are very high between parameters because it does not include the information about the off-diagonal elements of FIM. The D-optimal criterion is the most widely used among these standard criteria because of its simple geometrical interpretation and the invariant nature to any dimensional transformation applied on the model parameters. However, it gives no favors to any parameters of more importance, and also the correlation is not well considered between parameter pairs. The E-optimal design is to minimise the largest error of the covariance matrix while in other directions estimation errors could still be large. The modified E-optimal design is to reduce the conditional number of FIM which is to make the confidence ellipsoid as spherical as possible, but the whole volume of the ellipsoid may turn out to be large which indicates limited information content. In addition, the criterion in the modified E-optimal design is discontinuous due to its mathematical properties and may cause convergence problems in gradient-based optimisation. There are also other design criteria (such as G-, L-, C-optimality) proposed for OED which are less used in literature. A detailed description can be found in (Ljung, 1998).

Earlier work has suggested that no single design criterion can be applicable for all design problems or in all systems (Yue et al., 2013). The OED results are different when different design criteria are used. For a given system, one particular design criterion may be superior to others; but it does not mean this criterion plays well in other designs. Therefore, it is recommended that different scalar metrics should be tried and compared in the OED. Those

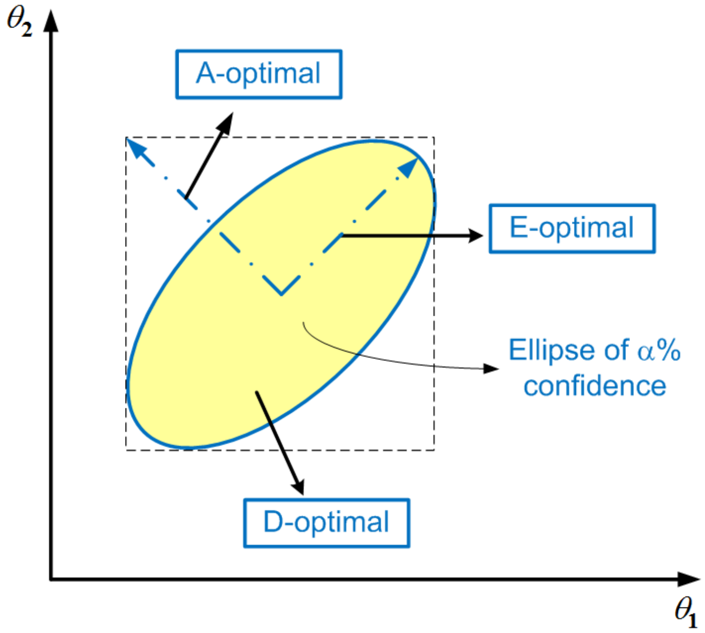


Fig. 3.4 Geometrical interpretation of the standard criteria for OED

standard design criteria might not meet the requirements for complex OED tasks, such as multiple objectives, constraints on parameter correlations and others. Thus, modification of these ‘alphabetic’ design criteria is a research area of recent interests space (Zen and Tsai, 2004).

3.3.3 Experimental design for input and observation factors

In OED for most biochemical systems, it is mainly focused on the manipulation of input variables, choice of sampling time points and/or selection of measurement set from available measurable output variables. The variation of input variables will change the system responses and in some case, e.g. fed-batch process, the input variables are time dependent. In the design of observations, such as sampling profile and measurement sets, the dynamic responses will not change under given input manipulations. Design for input factors and observation factors are therefore quite different. Three typical OED problems for biochemical systems have been discussed in the following.

(1) *OED of input factors*

One typical design problem in biochemical systems is the input design of a fed-batch process, such as the fed-batch bioreactor of fermentation processes (Berkholz et al., 2000). In this case, the input factors are time-varying continuous variables making the design problem an infinite dimensional optimisation problem, which is very difficult to solve. One approach is to transfer the original problem into a relaxed finite dimensional non-linear programming (NLP) dynamic optimisation problem by approximating the time-varying input variables with a discrete form of inputs. The most widely used direct dynamic optimisation methods are the sequential methods and the simultaneous methods (Biegler et al., 2002). In the simultaneous methods, for example the complete parametrisation method, all the states and control variables are discretised which leads to large-scale non-linear programming dynamic problems. As a result, the optimisation problem is directly coupled with the solution of a dynamic system, where the solution of the system is solved only once at the optimal point, therefore can avoid intermediate solutions and reduce the computational efforts. Efficient methods are required to solve large NLP problems and appropriate placement of finite elements need to be considered in order to maintain the accuracy of the discretised models. Compared with the simultaneous methods, the sequential methods, also called control vector parametrisation methods (CVP) (Vassiliadis et al., 1994a,b), are more popular methods that have been widely used in various applications. With sequential methods, the control variables are discretised which are solved iteratively. Given a dynamic model, the whole duration of an experiment is divided into several time intervals. Within each interval the control variables are discretised by low order polynomial functions of time, and therefore, the optimisation is performed with respect to the corresponding polynomial coefficients (Balsa-Canto et al., 2008). Note that the selection of parametrisation of the input or control variables is problem dependent. It is important to consider which functions, step-wise, step-linear, quadratic, etc., should be used to approximate the stimuli.

Figure 3.5 illustrates how the time-varying input signal is manipulated in a fed-batch reaction system by using CVP methods. It can be seen that the number of intervals and the switching points, as well as the level of the input within each interval need to be designed simultaneously. During the OED, the system responses need to be calculated again each time the input values are changed.

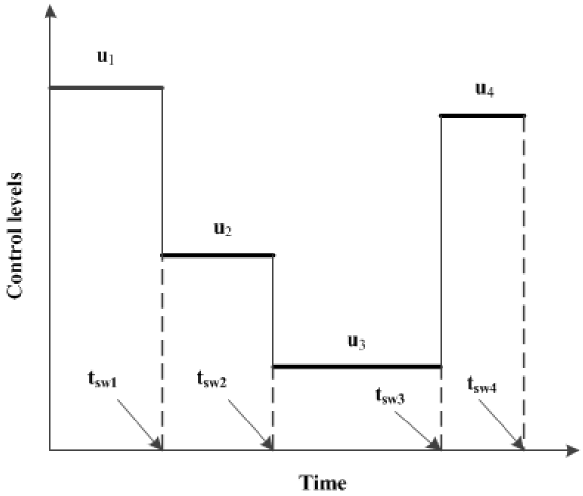


Fig. 3.5 An illustration of the CVP technique for the piecewise constant case over four intervals

The input experimental design has been applied in various applications. The OED techniques to a MAPK pathway system is applied in order to find the optimal input profile of Mek (Faller et al., 2003), in which a polynomial parametrisation is used for the input function. The feeding strategy of the *Trichosporon cutaneum* system has been investigated in (Baltes et al., 1994). It is concluded that fed-batch processes with small rates of change in substrate concentration allow a reasonable estimation of kinetic parameters. Similar results have been obtained in (Lindner and Hitzmann, 2006) from the input design of a Michaelis-Menten enzyme kinetic process. It is demonstrated that an enzyme feed will not improve the estimation quality but that substrate feedings with small volume flow could significantly increase parameter estimation precision.

(2) *Sampling time design*

The optimal time selection for parameter estimation in a dynamic signal pathway model is investigated (Kutalik et al., 2004). Instead of transferring the design problem to a continuous optimisation problem, the Powell's conjugate direction method (Powell, 1964) is applied to discrete optimisation. The main advantage of this method is that it is computationally efficient and it is easy to find a solution that is close to the optimal sampling. A first-order model for a drug distribution process has been developed to determine the optimal sampling time for measurement so as to reduce the estimation uncertainty of the pharmacokinetic parameters (Asyali, 2010). Two process parameters are added into the objective function, which are the maximum concentration of the drug and the time to reach the maximum concentration. This combines the parameter identification with process optimisation. From the numerical viewpoint, a spline approximation instead of the traditional stepwise linear approximation is used and the minimum number of sampling times and the corresponding locations is determined. The OED for a Michaelis-Menten kinetic model has been investigated in order to show that the optimal sampling strategy has advantages over the equidistant measurement points in parameter estimation (Ataíde and Hitzmann, 2009). It is suggested the OED should be used when the rough estimated parameters are near the true values, while equidistant measurements are better used when the predicted parameter values are not accurate.

The D-optimal sampling time design is compared to the equidistant sampling strategy for a cardinal parameter model in a food predictive microbiology system (Van Derlinden et al., 2013). For both methods a full factorial design and a Latin-square design plan were constructed and evaluated. It is found that for the design of a single factor such as temperature, the OED yields a better result than the equidistant design, but when considering multiple design factors which affect the system together, the equidistant level selection can obtain results with the similar quality as the optimal design. By using the Latin-square design to reduce the sampling points in experiments, the parameter estimation results have the same quality as the full factorial design. However, since the construction of experimental design is based on the prior information, e.g. most experimental design is based on the roughly estimated parameter values, reducing experiments by using Latin-square design may require much less prior information and make the estimation more sensitive to the experimental variations.

The sampling strategies are especially important in biochemical systems and systems biology where experimentations are most often time consuming and cost intensive.

(3) *Measurement set selection*

In most biochemical systems, not all reaction species in the process can be measured and the data information contained in different state variables can be quite different. For this reason, it is important to find the most valuable state variables, the measurement of which can provide more informative data than from other states. With measurement set design, another benefit is to find potentially important variables which are ignored in the current measurement settings. The robust measurement set design for an IkB-NF-kB signalling pathway network has been investigated in (Brown et al., 2008). The Maximin and the Bayesian RED are compared in the selection of measurement set and it is found that the Bayesian method is preferred in the case of large parameter uncertainties. The design of measurement sets for a signal transduction pathway model is also considered in (Jia and Yue, 2012). The effects of different measurement observables on the quality of measurement data are compared. Five best measurement states are determined from which the unknown parameters can be inferred with the best possible statistical quality.

In the above three OED problems, the first one is on the design of input manipulation; the second and third ones are regarding observation design. It should be noted that the input change would affect the system's dynamic response, while the observation design is developed based on available measurement data under given input conditions.

3.4 Numerical optimisation algorithms for solving OED problems

Once the optimisation formulation is proposed based on the OED targets, it is required to employ efficient algorithms to determine its optimal solutions. The numerical solvers can be roughly classified into two main groups: local and global methods. In general a local method is designed to generate a sequence of solutions that will converge to a local optimum that is close to its initial guess. The local optimisation algorithms are especially suited for convex optimisation problems. However, for the cases where suboptimal solutions may appear there is a high risk of getting trapped into a suboptimal solution. In this regard, optimisation techniques on non-convex OED optimisation problems are developed.

In cases where design variables are time dependent and continuous, it is required to approximate these design variables to reformulate the infinite dimensional problems as finite dimensional problems. Different optimisation methods have been proposed to achieve this goal. One early such experimental design problem was firstly transferred into an optimal control problem for dynamic models in (Espie and Macchietto, 1989). The control variables were parametrised as an user-defined number of piecewise continuous functions, and therefore makes a finite-dimensional optimisation problem. The authors also mentioned that this algorithm can be used on both model discrimination and parameter estimation. This method is similar to the CVP method, in which the experimental durations are divided into several time intervals, and within each interval time-dependent control variables are approximated by low-order Lagrange polynomials with a normalised time formula. The selection of order for this function is determined by the real experimental possibilities and experimenters' expertise. As a result, a non-linear programming problem is formulated and the optimisation is reduced to determine coefficients of polynomial functions. This method is particularly suitable for the OED applications and provides guidance for testing a variety of experimental conditions. Apart from the CVP methods, the complete parametrisation methods which belong to the simultaneous approaches are also widely used. The complete parametrisation methods are proposed to discretise both the control variables and the model states which result in large scale non-linear programming problems. The main advantage of this kind of method is that it can avoid repeated integration of DAE models. However, specific optimisation methods are required to get the optimal solution for these simultaneous approaches.

For OED problems with a large number of design variables or variables with large degrees of freedom (e.g. determine ten best samplings for an experiment or design an input with ten feeds), the local methods usually can only find the suboptimal solution due to the multi-modality of the optimisation problem. In some cases the standard gradient-based solvers even fail to converge to a local solution because of the non-smooth and highly non-linear nature of OED problems. Several algorithms based on stochastic methods have been applied to find the global optimal solution. In the work of (Banga et al., 2002), the numerical aspect for OED problems are considered. The authors point out that the cost functions are most often non-convex with the increased complexity of dynamic models as well as sophisticated design targets; traditional gradient-based methods may result in local solutions which largely depend on the initial guesses of the design factors. Therefore, stochastic methods of global optimisation to solve the OED problem are suggested. The integrated controlled random search algorithm and the differential evolution method are applied for solving an OED problem of a fed-batch bioreactor. It is demonstrated that stochastic methods show better performance than local optimisation algorithms on the ability to find the global optimum (Banga et al., 2002). Similarly, the numerical aspects connected with the experimental design methods for biochemical pathways are investigated (Rodriguez-Fernandez et al., 2006). The local optimisation method and the hybrid method for the OED of a biochemical pathway model are compared. Hybrid methods are, as the name suggested, combining the advantages of local methods with that of global methods. A global method is firstly applied to identify the promising regions of the searching space because it can explore the whole parameter space. The local method is then used to determine the minimum within the specified space owing to its fast convergent nature. The simulation result shows that the computational time is acceptable and the robustness of the solution is guaranteed. It is stated that applying the hybrid optimisation method can keep advantages of both local and global optimisation algorithms.

Although most stochastic algorithms cannot guarantee global optimality with certainty, the fact that they can get a lower bound for the objective function makes them an appealing choice than the traditional gradient-based optimisation methods. In this work, two population based methods, the PSO algorithm and the differential evolutionary algorithm will be employed to solve the non-convex input design problem. The candidate solutions are first generated by random choice of the initial design factors. Then these solutions will be updated and eventually converge to the global optimum by using some mechanisms inspired by the biological evolution: reproduction, recombination and selection.

These discussions are mainly about non-convex optimisation problems. For convex optimisation problems, linear or nonlinear programming methods can be applied to obtain the global optimal solution.

3.5 Summary

In this chapter, the basic theories for model-based OED are described for model discrimination and parameter estimation. Recent developments and applications of OED techniques ofr biochemical systems are reviewed. The numerical methods to solve OED problems are discussed. This chapter provides the basics for further model-based OED development in this thesis.

Chapter 4

Dynamic Modelling of Two Biochemical Systems and Analysis

This chapter provides a detailed description of two biochemical systems. The first one is a new hypothetical model which is proposed to represent a typical class of kinetically controlled synthesis processes. A lot of chemical processes follow a similar kinetic reaction scheme. Therefore, model identification of this kind of system is of particular importance as it can facilitate understanding of system dynamic behaviour and improvement of process optimisation. The ODEs model to express the enzyme reaction system is first proposed. By solving the model equations, the system behaviour and the connection between model parameters and state variables are given. Next, some preliminary analysis will be conducted to investigate the system's equilibrium state and how model parameters affect the system behaviour. In particular, the identifiability of this model will be analysed in order to find out whether this kinetic model is structurally identifiable and which parameters can be identified using practical experimentation. The second one is a lab-scale biochemical system which describes the production process of biodiesel by using waste vegetable oil (Price et al., 2014a,b). Green chemistry has attracted increasing interest in recent industry where the main aim is to eliminate hazardous chemicals, minimise waste generation and energy consumption. The biocatalyst for the biodiesel production process is becoming an interesting research field. However, the main drawbacks to the use of enzymes are the usually low product yields and poor enzyme stability. Therefore the investigation of this system to facilitate further analysis and process improvement is important.

The enzyme reaction model and the enzymatic biodiesel production system model will be used in later experimental design to show how model-based OED techniques can be applied in biochemical processes to facilitate model parameter identification.

4.1 Establishment of a benchmark model - Enzyme reaction system

Enzymatic processes exist in quite a number of chemical, biochemical and biological systems, the understanding of which is important for system prediction and process control. In this section, a new enzyme reaction model is proposed which can represent a typical class of kinetically controlled synthesis processes. The proposed model contains important kinetic reaction features and moderate complexity. This model can be used as a benchmark problem for development of OED strategies.

Figure 4.1 illustrates an enzymatic process with kinetically controlled synthesis which will be used as a test model throughout the whole research work (Yue et al., 2013). The reactions taking place simultaneously during the kinetically controlled synthesis can be distinguished in this scheme. The initial reactants are donor substrate (S), nucleophile (N), and the enzyme catalyst (E) (red fonts). The final product is expressed as Q and the side product R (green ones). All the reactants are assumed to be perfectly mixed in the reactor. Firstly the substrate S binds to the enzyme E to form the enzyme-substrate complex ES, and then ES can be decomposed into another compound E^* and the leaving group product P. E^* can either react with the nucleophile N to form EQ or be hydrolysed to produce ER. EQ can be decomposed into the required product Q and enzyme E, while ER can be decomposed into the hydrolysis by-product R and enzyme E. During the whole reaction process, all the reactions are reversible, except for the decomposition of ER to give E and R. Due to the characteristic of enzyme, it only catalyses the reactions and at the end of the reaction, the amount of enzyme remains the same as before the chemical reactions start.

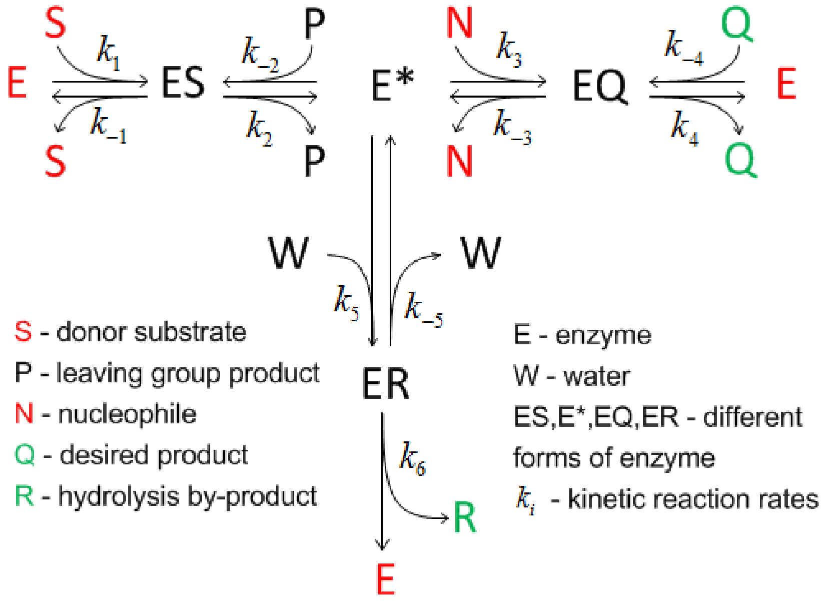


Fig. 4.1 Reaction mechanisms of enzyme reaction system

A number of enzymatic processes have a similar reaction scheme which is known as kinetically controlled synthesis. In this system, the desired product Q is not thermodynamically most favourable one. The hydrolysis by-product R will dominate at long times. The kinetic reaction scheme proceeds as shown in Table 4.1.

The parameters k_1, k_2 etc. describe the reaction rates of the system. Among those reaction species, Q, S, P, N and R are measurable species in experiments. It is difficult to measure different forms of enzymes due to their low concentrations and limited measurement techniques. A chemical reaction is always affected by the surrounding environment and

Table 4.1 Kinetic mechanism for the enzyme reaction system

Reactions		Rate of reactions
$E+S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES$	The substrate binds to the enzyme and then releases intermediate compound E^*	$k_1E\cdot S-k_{-1}ES$
$ES \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} P+E^*$		$k_2ES-k_{-2}P\cdot E^*$
$N+E^* \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EQ$	The E^* then reacts with nucleophile to generate the desired product Q	$k_3N\cdot E^*-k_{-3}EQ$
$EQ \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} Q+E$		$k_4EQ-k_{-4}Q\cdot E$
$E^*+W \underset{k_{-5}}{\overset{k_5}{\rightleftharpoons}} ER$	The enzyme complex E^* can react with water to produce ER and then release the side product R irreversibly.	$k_5W\cdot E^*-k_{-5}ER$
$ER \xrightarrow{k_6} E+R$		k_6ER

factors such as the inactivation of enzyme, reactant instability, effects of pH and temperature, etc. In order to conveniently investigate the system from the perspective of experimental design, all these complications have been removed in this test system. All reactants will be added to the reactor before the start of the reaction therefore the volume of this system is treated as constant. There are 10 reaction species and 11 parameters in the model. Following the mass action principles, the enzyme reaction system can be expressed as 10 ODEs:

$$
\begin{aligned}
\frac{dE}{dt} &= -k_1 \cdot E \cdot S + k_{-1} \cdot ES + k_4 \cdot EQ - k_{-4} \cdot E \cdot Q + k_6 \cdot ER \\
\frac{dES}{dt} &= k_1 \cdot E \cdot S - k_{-1} \cdot ES - k_2 \cdot ES + k_{-2} \cdot E^* \cdot P \\
\frac{dE^*}{dt} &= k_2 \cdot ES - k_{-2} \cdot E^* \cdot P - k_3 \cdot E^* \cdot N + k_{-3} \cdot EQ - k_5 \cdot W \cdot E^* + k_{-5} \cdot ER \\
\frac{dEQ}{dt} &= k_3 \cdot E^* \cdot N - k_{-3} \cdot EQ - k_4 \cdot EQ + k_{-4} \cdot E \cdot Q \\
\frac{dER}{dt} &= k_5 \cdot W \cdot E^* - k_{-5} \cdot ER - k_6 \cdot ER \\
\frac{dS}{dt} &= -k_1 \cdot E \cdot S + k_{-1} \cdot ES \\
\frac{dP}{dt} &= k_2 \cdot ES - k_{-2} \cdot E^* \cdot P \\
\frac{dN}{dt} &= -k_3 \cdot E^* \cdot N + k_{-3} \cdot EQ \\
\frac{dQ}{dt} &= k_4 \cdot EQ - k_{-4} \cdot E \cdot Q \\
\frac{dR}{dt} &= k_6 \cdot ER
\end{aligned}
\tag{4.1}
$$

Assuming the left hand side of (4.1) to be zero, the following conservation constraints can be obtained by simple algebraic calculation:

$$
\begin{aligned}
& ES + S + P = C_1 \\
& EQ + N + Q = C_2 \\
& E + ES + EQ + E^* + ER = C_3 \\
& ES + EQ + S + Q + R + E^* + ER = C_4
\end{aligned}
\tag{4.2}
$$

The above constraints can also be deduced using chemical logics, based on the principle of conservation of atoms. From the reaction scheme shown in Figure 4.1, it can be seen that *S* only reacts with *ES* to generate *P* or reversibly from *P* to *S*, so the total amount of

$S+ES+P$ is constant. A similar case happens among N, EQ and Q, therefore, the total amount of $N+EQ+Q$ is also constant. As well known that enzyme only catalyses reactants and it will finally all go back to itself. Therefore, the total amount of enzyme complexes plus the enzyme is also constant. However, the last constraint in (4.2) is not easily seen from the figure, but it does come from chemical logics as conservation of the acyl portion of S (the part that remains bound to the enzyme in E^*). It should be noted that all the constraints in (4.2) apply for the whole reaction process. C_1, C_2, C_3 and C_4 are constants which can be determined by the initial conditions of the related variables. Before any reaction of this biochemical system, only the enzyme (E), the substrate (S) and the nucleophile (N) are added as the initial reactants. All the other state variables, P, Q, R, ES, E^*, EQ and ER are equal to zero. So we have

$$C_1 = C_4 = S_0,\ C_2 = N_0,\ C_3 = E_0$$

S_0, N_0 and E_0 are initial conditions of S, N and E, respectively. Since the concentrations of enzyme and enzyme related compounds are much smaller than that of other species, the conservation constraints can be simply expressed as

$$S+P \approx S_0,\ S+Q+R \approx S_0,\ N+Q \approx N_0$$

According to the above conservation analysis, we can find that among the ten reaction species, only six are independent variables. The other four variables can be formulated as algebraic equations. Therefore the ten state variables can be classified into two groups: the independent variables and the dependent variables. Here we choose $[S, P, N, Q, R, E^*]$ as the six independent variables and the other four $[E, ES, EQ, ER]$ as the dependent variables. Therefore the original ODEs can be reduced into 6 ODEs and 4 algebraic equations, shown as follows:

$$
\begin{aligned}
&\frac{dE^*}{dt} = -(k_5W + k_{-5})E^* - k_2S + (k_{-5} - k_2)P + (k_{-5} - k_{-3})N \\
&-k_{-3}Q - k_{-5}R - k_{-2}E^*P - k_3E^*N + k_2S_0 + (k_{-3} - k_{-5})N_0 \\
&\frac{dS}{dt} = (-k_1E_0 + k_1S_0 - k_{-1})S - k_{-1}P - k_1(S + Q + R)S + k_{-1}S_0 \\
&\frac{dP}{dt} = -k_2(S + P) - k_{-2}E^*P + k_2S_0 \\
&\frac{dN}{dt} = -k_{-3}(N + Q) - k_3E^*N + k_{-3}N_0 \\
&\frac{dQ}{dt} = -k_4(N + Q) - k_{-4}(S + Q + R)Q + k_4N_0 - k_{-4}(E_0 - S_0)Q \\
&\frac{dR}{dt} = k_6(-E^* + P + N - R) - k_6N_0 \\
&ES = S_0 - S - P \\
&EQ = N_0 - N - Q \\
&ER = -N_0 + N + P - E^* - R \\
&E = E_0 - S_0 + S + Q + R
\end{aligned}
\tag{4.3}
$$

In this reaction system, the amount of water, W, is very large and can be seen as constant. Therefore, in further analysis k_5W is used as a parameter instead of k_5 alone. An arbitrary upper bound for N_0 and S_0 is 1 $mol \cdot L^{-1}$ and a further constraint for E_0 is $S_0/E_0 \geq 10^3$ for sensible operation so the upper bound for E_0 is $10^{-3} mol \cdot L^{-1}$. The nominal condition of the initial input values is listed in Table 4.2:

From the empirical experience it is known that the second order rate constants, k_1, k_{-2}, k_3 and k_{-4} never exceed $10^9 L \cdot mol^{-1} \cdot s^{-1}$. However, there are not clear theoretical upper limits for the other parameters. These parameter values are usually determined from experimental data. For a feasible process, the rate of Q formation should be at least no less than that of R. However, if R always shows a much slower formation rate than Q formulation, the side reaction becomes insignificant and results in no complications of the kinetically controlled

Table 4.2 A list of initial values for state variables in the enzyme reaction system

State variables	Reactants	Initial values ($mol \cdot L^{-1}$)
x_1	E^*	0
x_2	S	0.8
x_3	P	0
x_4	N	0.9
x_5	Q	0
x_6	R	0
x_7	E	1.5e-5
x_8	ES	0
x_9	EQ	0
x_{10}	ER	0

process. In order to investigate the characteristics of Q formulation in the long term, the parameters are set in a way that in the early stage the rates of Q formulation and R formulation are comparable, the value of which are given in Table 4.3.

Table 4.3 A list of nominal values for kinetic parameters in the enzyme reaction system

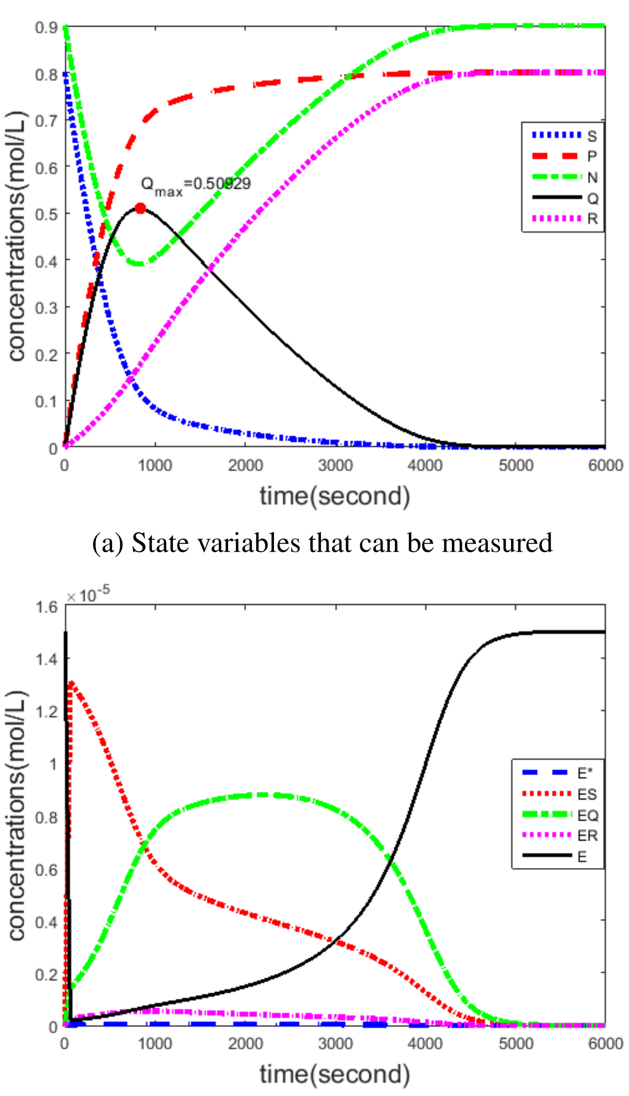
Kinetic parameters	Nominal values	Units
k_1	1e5	$L \cdot mol^{-1} \cdot s^{-1}$
k_{-1}	1e3	s^{-1}
k_2	100	s^{-1}
k_{-2}	1e4	$L \cdot mol^{-1} \cdot s^{-1}$
k_3	5e4	$L \cdot mol^{-1} \cdot s^{-1}$
k_{-3}	200	s^{-1}
k_4	1e3	s^{-1}
k_{-4}	2e4	$L \cdot mol^{-1} \cdot s^{-1}$
k_5	5e3	s^{-1}
$k_{-5}W$	100	s^{-1}
k_6	500	s^{-1}

The mathematical model of this enzymatic reaction system is thus proposed with rough parameter values and initial input conditions. It covers a typical kinetic reaction scheme with moderate model complexity. The effect of the side reaction on the generation of the desired product is considered. The difficulty in the identification of kinetic parameters is also reflected in this model.

4.2 System analysis for enzyme reaction system

4.2.1 Equilibrium state analysis

Figure 4.2 demonstrates the concentration time profiles of all state variables of the enzyme reaction system. Nominal parameter values and initial input conditions are used in this simulation (Table 4.2 and Table 4.3) which is called the nominal simulation condition in later work. It can be seen that the concentrations of all the enzyme complexes, the substrate and the desired product are equal to zero ($ES(\infty) = E^*(\infty) = EQ(\infty) = ER(\infty) = S(\infty) = Q(\infty) = 0$) at the equilibrium state, and the values of other four state variables, P, E, N and R, at the equilibrium state, are dependent on the initial conditions via the conservation laws. In this case, $P(\infty) = R(\infty) = S_0$, $N(\infty) = N_0$, and $E(\infty) = E_0$. This result can also be easily calculated by taking all the ODEs in (4.1) to be zeros. From the biochemical viewpoint, the enzyme acts only as catalyst and it will all return to its free form E at the equilibrium state, which brings $E(\infty) = E_0$, whereas all other enzyme complexes will be zero. Also, in this system only the generation of the hydrolysis product R is irreversible. Therefore, at the end of the reaction all the desired product Q and the substrate S will be exhausted and converted to the leaving group product P and the hydrolysis by-product R. The output of major interest is the product Q, which increases during the early stage of the reaction and then decreases until reaching a zero equilibrium state. The production of Q is a balance of kinetic and equilibrium effects. The equilibrium position will always be essentially all R due to the irreversible reaction to generate R. The production of R is completely irreversible, therefore at the equilibrium state, all S and Q will be converted to R. The kinetic constants are such that R is not produced quickly. In the early stage, E^* will react much more rapidly with N than with W. Hence substantial amounts of Q are formed in the early stage. Once S is depleted, the reversible reactions leading to Q start to go in reverse. Much of the E^* accumulated still reacts with N to go back to Q, for no net change. But the small fraction of



(a) State variables that can be measured

(b) Free Enzyme and enzyme complexes

Fig. 4.2 Time profiles for concentration of all state variables in enzyme reaction system

E^* that does react with W will go on irreversibly to R. Therefore, overall there is a slow but continuing conversion of Q, formed earlier, to R.

4.2.2 Preliminary analysis for process optimisation

From the perspective of process optimisation, the product of interest is Q, of which the production needs to be maximised. The kinetic reaction should be stopped once the production of Q achieves its maximum value. The substrate is considered to be expensive in enzyme reaction systems. By considering the total substrate added, this process can be treated as the optimisation of Q_{max}/S_0 and at the same time minimising the time to reach its peak value. Similar objective functions can be obtained for other input conditions, i.e. Q_{max}/N_0 and Q_{max}/E_0. For this non-linear system, it is difficult to obtain the analytic function for Q. Therefore it is necessary to find the effects of different control variables on the level of Q through numerical study.

(a) *Impact of S_0 to Q_{max}/S_0*

First we set E_0, N_0 and all the parameters at their nominal values; the substrate S varies from 0.01 to 1 $mol \cdot L^{-1}$ with the sampling interval of 0.01 $mol \cdot L^{-1}$; the relationship between the initial input condition and Q_{max} can be found which is shown in Figure 4.3. In Figure 4.3a, it can be seen that Q_{max}/S_0 is decreased with the increase of S_0 which means that with fixed values of N_0 and E_0, S_0 should be set to a small value so that the ratio between Q_{max} and S_0 can remain large. Figure 4.3b shows that with the increase of S_0, it needs more time to reach Q_{max}. Therefore, the input variable S_0 should be set to a small value which can make the time of reaching Q_{max} not too long and at the same time keep a high ratio of Q_{max}/S_0.

(b) *Impact of N_0 to Q_{max}/N_0*

A similar observation can be found between Q_{max} and N_0 (N_0 varies from 0.01 to 1 $mol \cdot L^{-1}$ with the step size of 0.01 $mol \cdot L^{-1}$) while S_0 and E_0 are kept constant (see Figure 4.3c).

(c) *Impact of* E_0 *to* Q_{max}/E_0

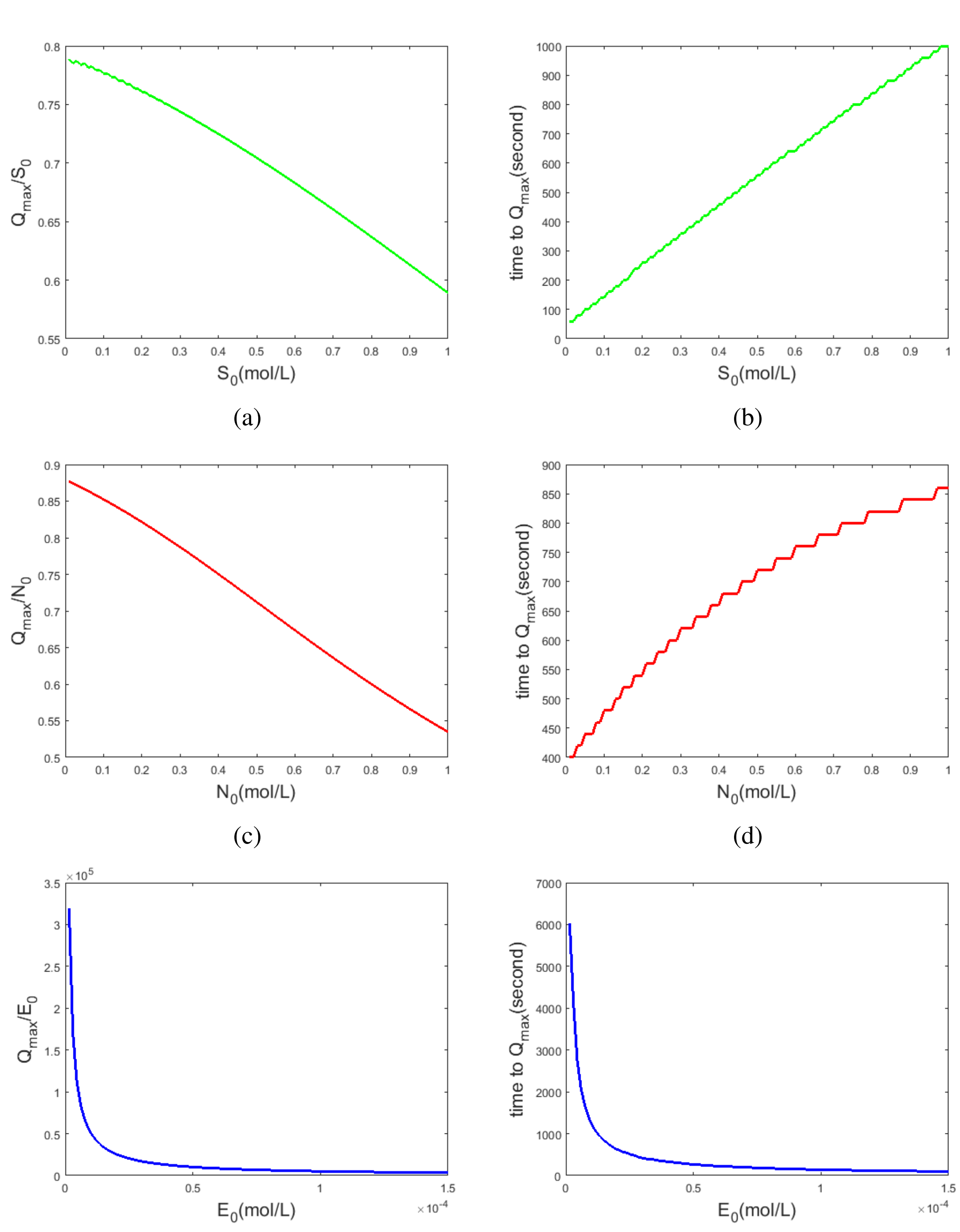
When we look at the influence of E_0 to Q_{max} (E_0 varies from 1.5e-6 to 1.5e-4 $mol \cdot L^{-1}$ with the step size of 1.5e-6 $mol \cdot L^{-1}$), it is found that the amount of E_0 should be set to a value that can catalyse all substrate to make sure that reaction is fast enough.

However, one should note that the reactants S and N should not be set to extremely small values because this will cause a waste of time and resources. Practically more conditions need to be taken into account such as the starting up cost of experiments and the prices of substrates and catalysts. Also, the three reactants, S, E and N, are interconnected with each other during the reaction process. When a small amount of substrate (e.g. 0.01 $mol \cdot L^{-1}$) is added, it will be inadequate to take the catalyst effect therefore making a waste of resources. In practical operations, one would like to add a large amount of reactants in a fixed volume of reactor to ensure a high yield of the desired product, and at the same time keep an appropriate proportion between the substrate and the enzyme to make sure the reaction is fast enough. For this purpose, at the initial stage the substrate should be set to a high value with a suitable amount of enzyme to catalyse the reaction.

4.2.3 Local sensitivity analysis

As discussed earlier in Section 2.4.1, parameter local sensitivities can be obtained by solving equations (2.15). The time profile of parameter local sensitivities to different state variables at the nominal simulation condition are shown in the following figures.

An immediate impression of kinetic parameter sensitivities from Figure 4.4 is that k_2, k_{-3} and k_{-5} are the three most important parameters during the whole reaction process. The



(a)

(b)

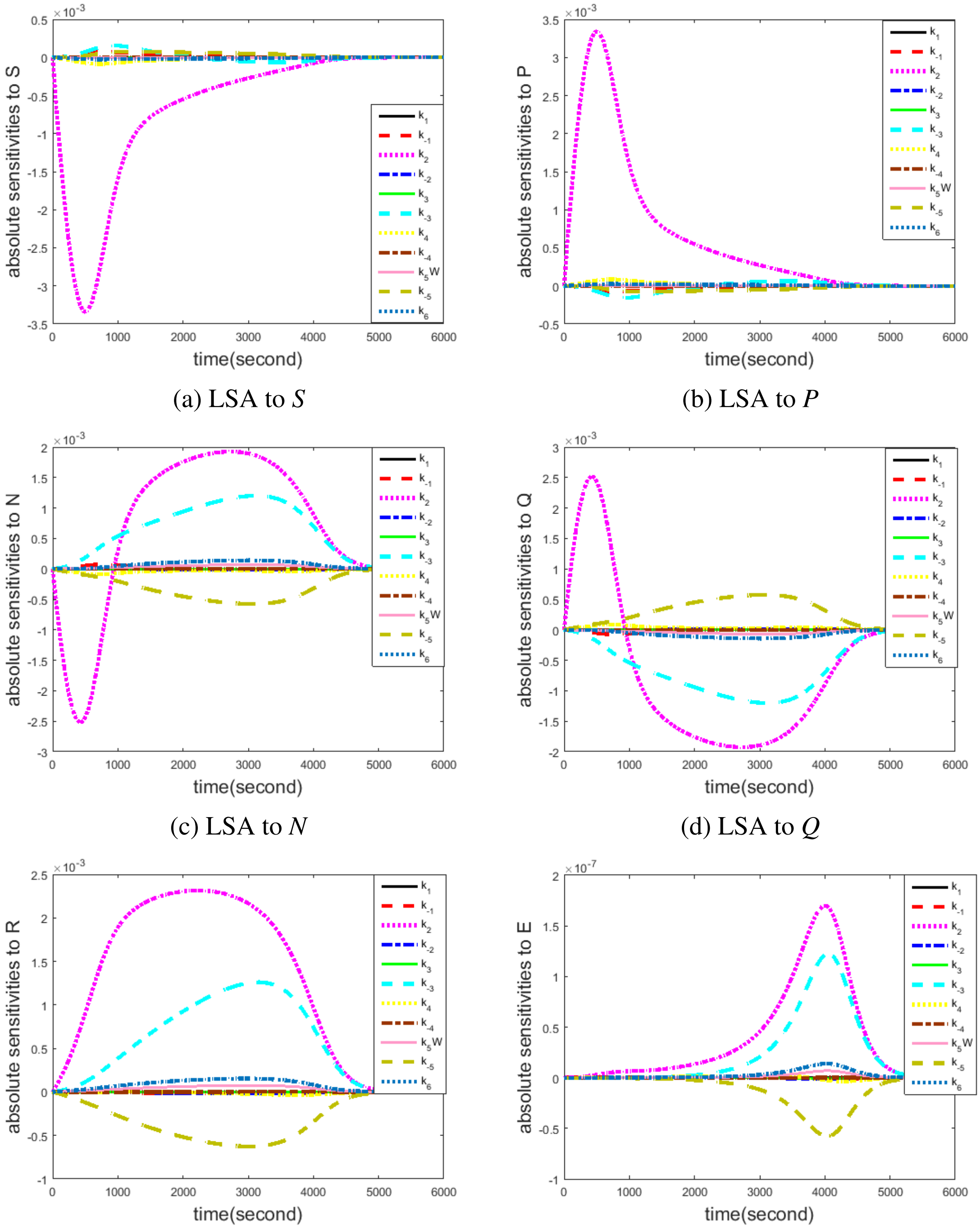
(c)

(d)

(e)

(f)

Fig. 4.3 Relationship between Q_{max} and initial input conditions



(a) LSA to *S*

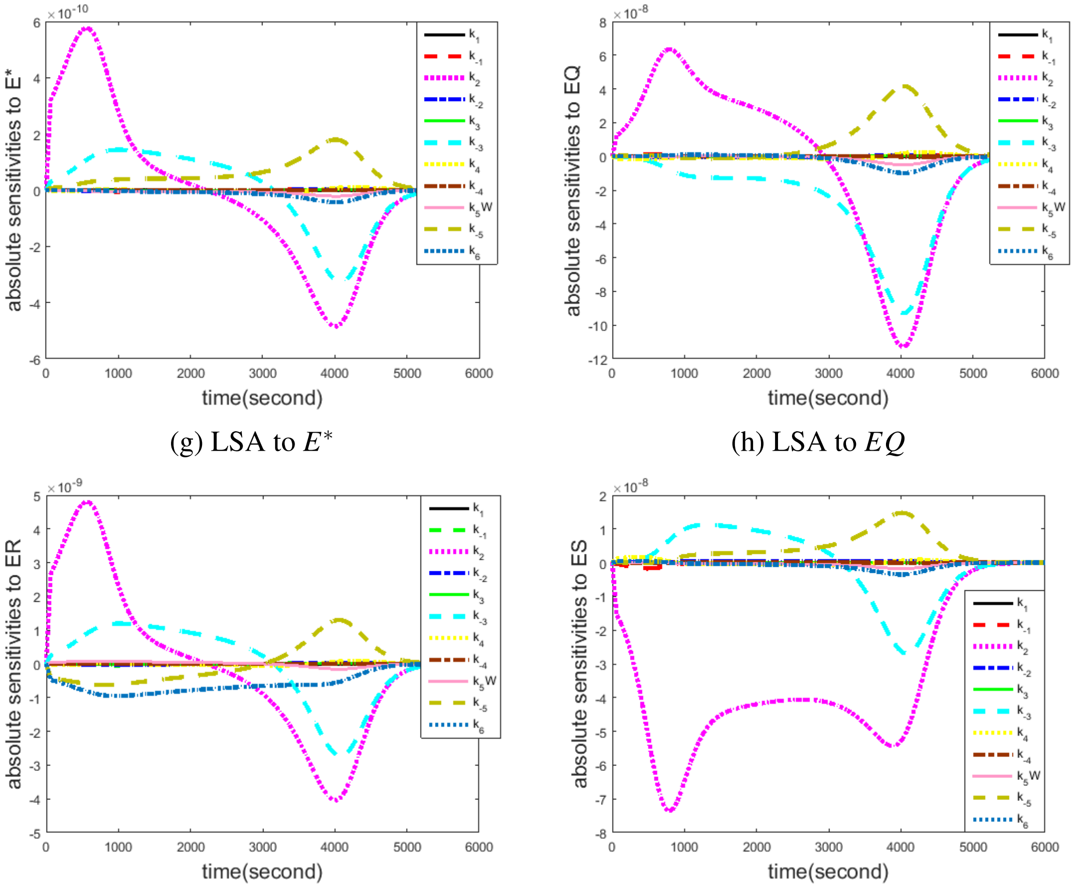
(b) LSA to *P*

(c) LSA to *N*

(d) LSA to *Q*

(e) LSA to *R*

(f) LSA to *E*



(g) LSA to E^*

(h) LSA to EQ

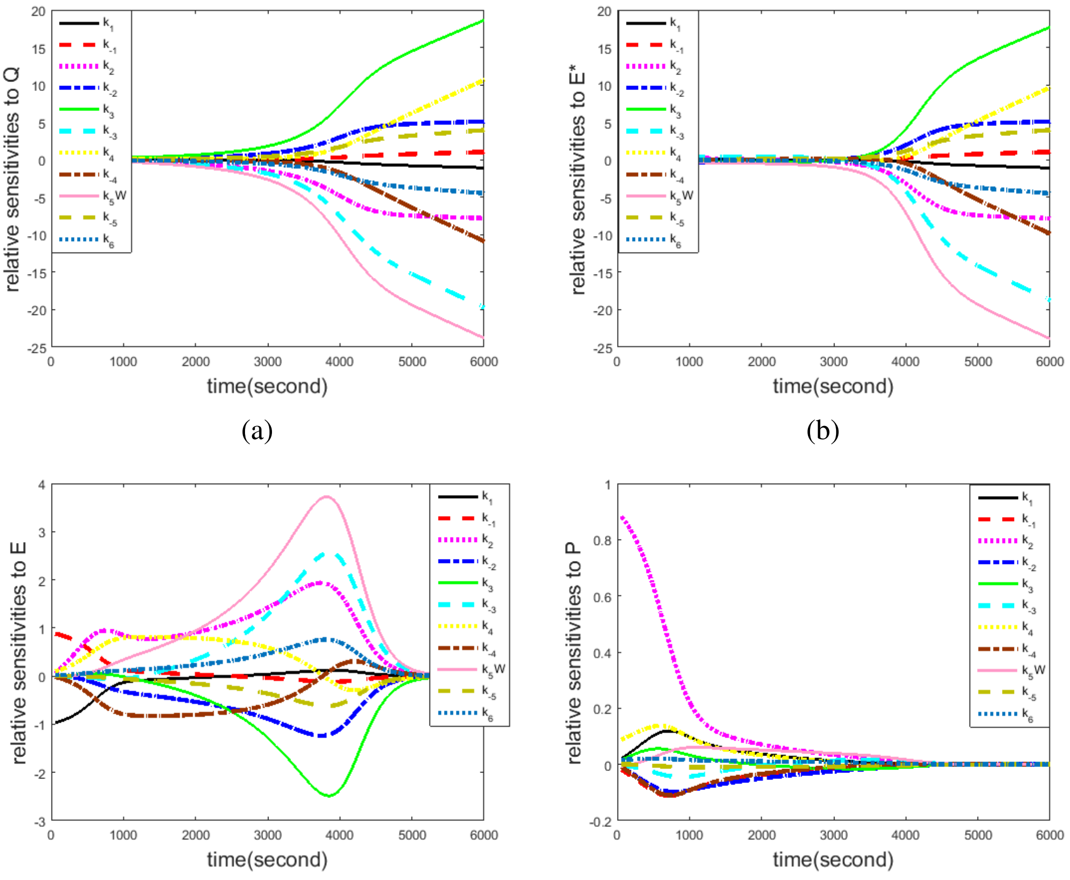
(i) LSA to ER

(j) LSA to ES

Fig. 4.4 Absolute parameter sensitivities to 10 state variables in the enzyme reaction system

averaged effect of kinetic parameters on those measurable state variables (S, P, N, Q, R) are much higher than that on enzyme complexes (E, E^*, ES, EQ, ER). Figure 4.4a and 4.4b shows that the influence of k_2 on state variables P and S is much bigger than that of other parameters, especially in the early stage of reaction where the generation of P and the depletion of S to form ES dominate the reaction process. Figure 4.4c, 4.4d and 4.4e demonstrate that the effect of the three important parameters is higher in the middle stage of the reaction process, when the production of R and the reversible reaction of Q start to dominate the process. However, Figure 4.4g -4.4j implies that parameter sensitivities to enzyme and its complex have significant changes at the early stage (around 800 seconds) and at the late stage of reaction (around the 4,000 seconds). At the start of the process most E binds to substrate to form different enzyme complexes, while at the late stage of reaction irreversible reaction to form R has dominated the process and all enzyme complexes are depleted and returned to E around that time. Kinetic parameters have different effects at different stages of the kinetic process. One should note that the difference of nominal parameter values can be up to several orders of magnitude among different kinetic parameters. The three most sensitive parameters, k_2, k_{-3} and k_{-5}, have smaller values, shown in Table 4.3. Similarly, the difference of the magnitude among different state variables are also of several orders. It can be found that the magnitudes of enzyme complexes are much smaller than that of the five measurable state variables. In order to allow direct comparison among different parameters and across different state variables, the relative sensitivities should be used to eliminate the effect of scale of parameters and state variables. Equation (2.16)-(2.19) will be applied for LSA and two different relative sensitivity analysis methods are compared in this case.

When using equation (2.17) for relative sensitivity analysis, there is a problem that the algorithm does not work when the states are zeros or very close to zeros. It can be seen from Figure 4.5 that the effect of kinetic parameters on Q and E^* have unreasonable large



(a)

(b)

(c)

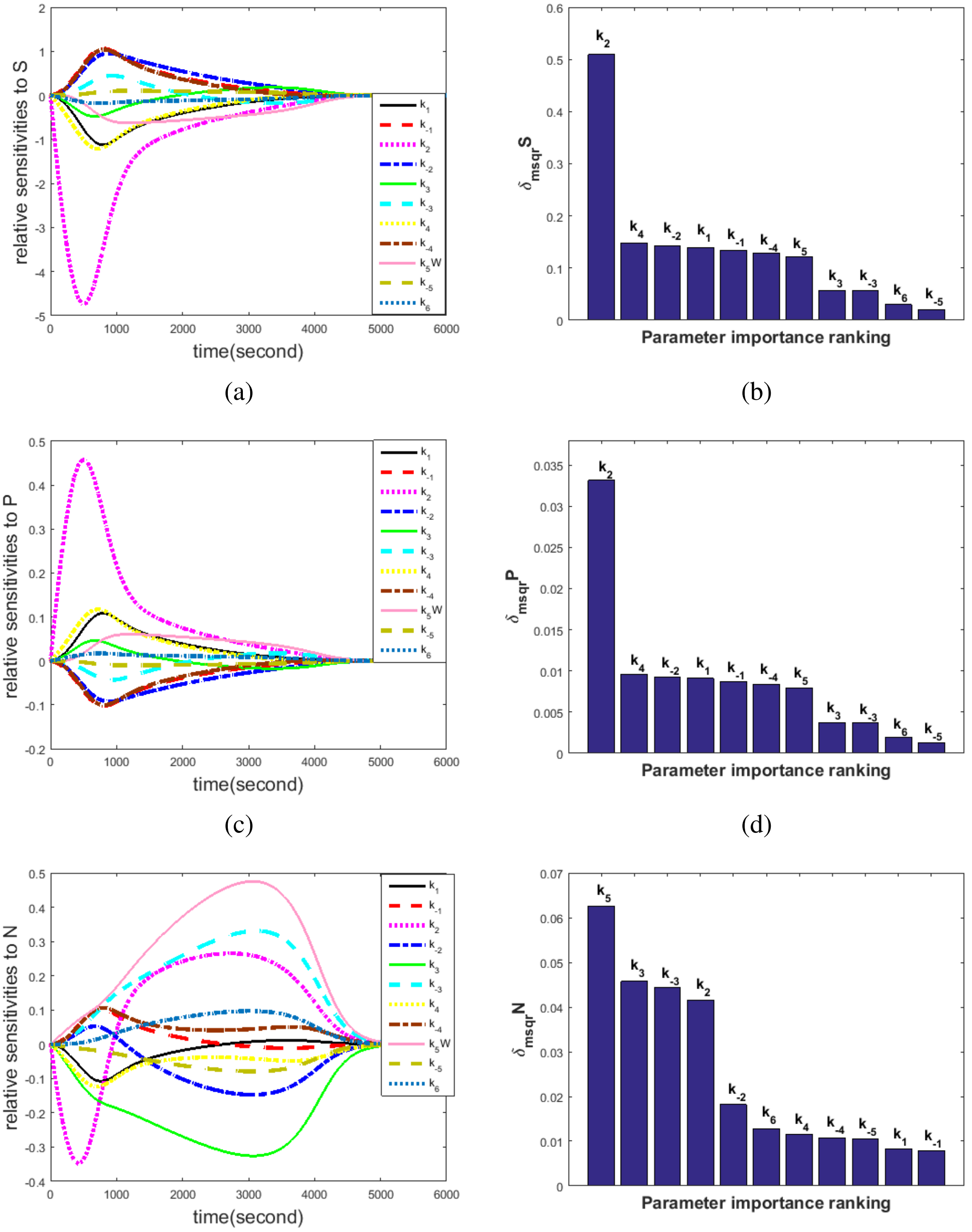
(d)

Fig. 4.5 Relative sensitivities to state variable Q, E^*, E, and P by using log function

sensitivities at the late stage of reaction, caused by roundoff or truncation errors in numerical treatment. At around 4,000 seconds, Q and E^* are exhausted, their concentrations keep reducing and tend to approach zero. Following equation (2.17), it will numerically result in large magnitudes of parameter sensitivities. The relative sensitivities of kinetic parameters to E and P seem reasonable by using the logarithmic function, shown in Figure 4.5c and 4.5d. Therefore, a proper time range should be chosen in order to avoid the numerical problem when using equation (2.17). Another option for relative sensitivity analysis is to use equation (2.16). The value of Δx_i is chosen to be the average value of each state variable across the whole experimental time and $\Delta\theta_j$ is the nominal value of each parameter.

The time profiles of relative sensitivities to different state variables are provided in the left column of Figure 4.6. It can be seen that the importance of kinetic parameters to different state variables are different from the results of absolute sensitivities. For those measurable state variables [S, P, N, Q, R], the kinetic parameter k_2 has a significant effect on them during the first 1,000 seconds of reaction except for R. For state variables S and P, the influences of other parameters are much less than that of k_2 at the early stage of reaction. It can also be found that all parameter effects on S and P are very low after around 4,000 seconds, because at this stage the substrate S is depleted and the reversible reaction from E^* to S is much less competitive than the reaction from E^* to R. Most parameters have more effect on state variables N, Q and R around 3,000 seconds, indicating their dominant roles in the middle stage of this kinetic reaction process, among which k_5W, k_3, k_{-3} and k_2 are the four most influential parameters. For those non-measurable state variables [E, E^*, ES, EQ, ER], it is obvious that the scaled parameter sensitivities are quite high around 4,000 seconds when all enzyme complexes tend to return to the free form E.

The averaged effect of parameter sensitivities along the whole reaction process are shown in the right column of Figure 4.6. it can be seen that the parameter importance ranking order for S is the same as for P, where k_2 has the largest effect. For state variable N, Q, and R,


relative sensitivities to S
time(second)
k$_1$
k$_{-1}$
k$_2$
k$_{-2}$
k$_3$
k$_{-3}$
k$_4$
k$_{-4}$
k$_5$W
k$_{-5}$
k$_6$
δ_{msqr}S
Parameter importance ranking
k$_2$ k$_4$ k$_{-2}$ k$_1$ k$_{-1}$ k$_{-4}$ k$_5$ k$_3$ k$_{-3}$ k$_6$ k$_{-5}$
relative sensitivities to P
δ_{msqr}P
k$_2$ k$_4$ k$_{-2}$ k$_1$ k$_{-1}$ k$_{-4}$ k$_5$ k$_3$ k$_{-3}$ k$_6$ k$_{-5}$
relative sensitivities to N
δ_{msqr}N
k$_5$ k$_3$ k$_{-3}$ k$_2$ k$_{-2}$ k$_6$ k$_4$ k$_{-4}$ k$_{-5}$ k$_1$ k$_{-1}$

(a)

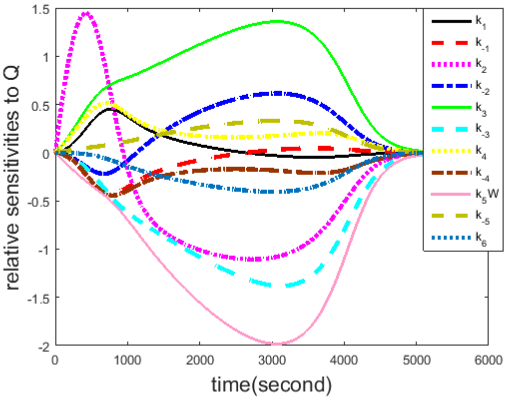
(b)

(c)

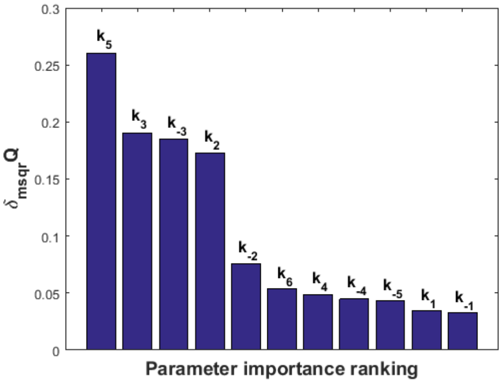
(d)

(e)

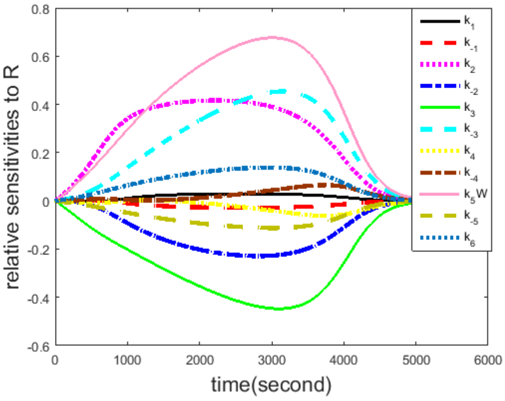
(f)


relative sensitivities to Q
time(second)
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
1000
2000
3000
4000
5000
6000
k_1
k_{-1}
k_2
k_{-2}
k_3
k_{-3}
k_4
k_{-4}
k_5W
k_{-5}
k_6

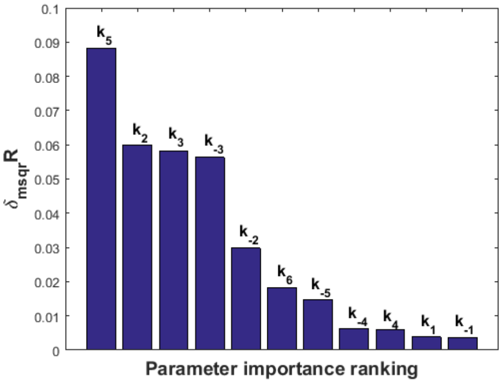
(g)


$\delta_{msqr}Q$
Parameter importance ranking
0.3
0.25
0.2
0.15
0.1
0.05
0
k_5
k_3
k_{-3}
k_2
k_{-2}
k_6
k_4
k_{-4}
k_{-5}
k_1
k_{-1}

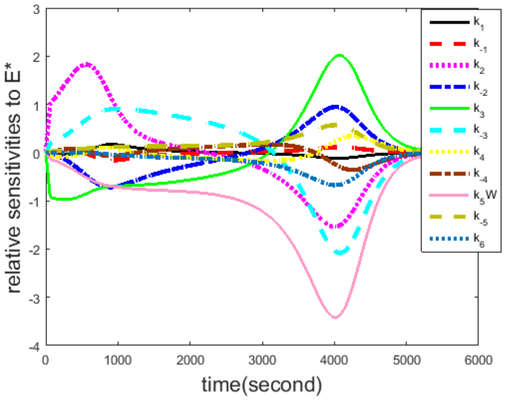
(h)


relative sensitivities to R
time(second)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0
1000
2000
3000
4000
5000
6000
k_1
k_{-1}
k_2
k_{-2}
k_3
k_{-3}
k_4
k_{-4}
k_5W
k_{-5}
k_6

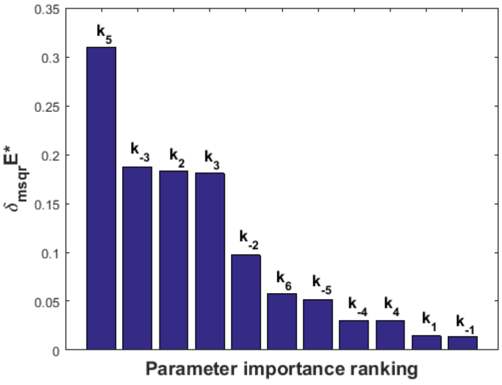
(i)


$\delta_{msqr}R$
Parameter importance ranking
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
k_5
k_2
k_3
k_{-3}
k_{-2}
k_6
k_{-5}
k_{-4}
k_4
k_1
k_{-1}

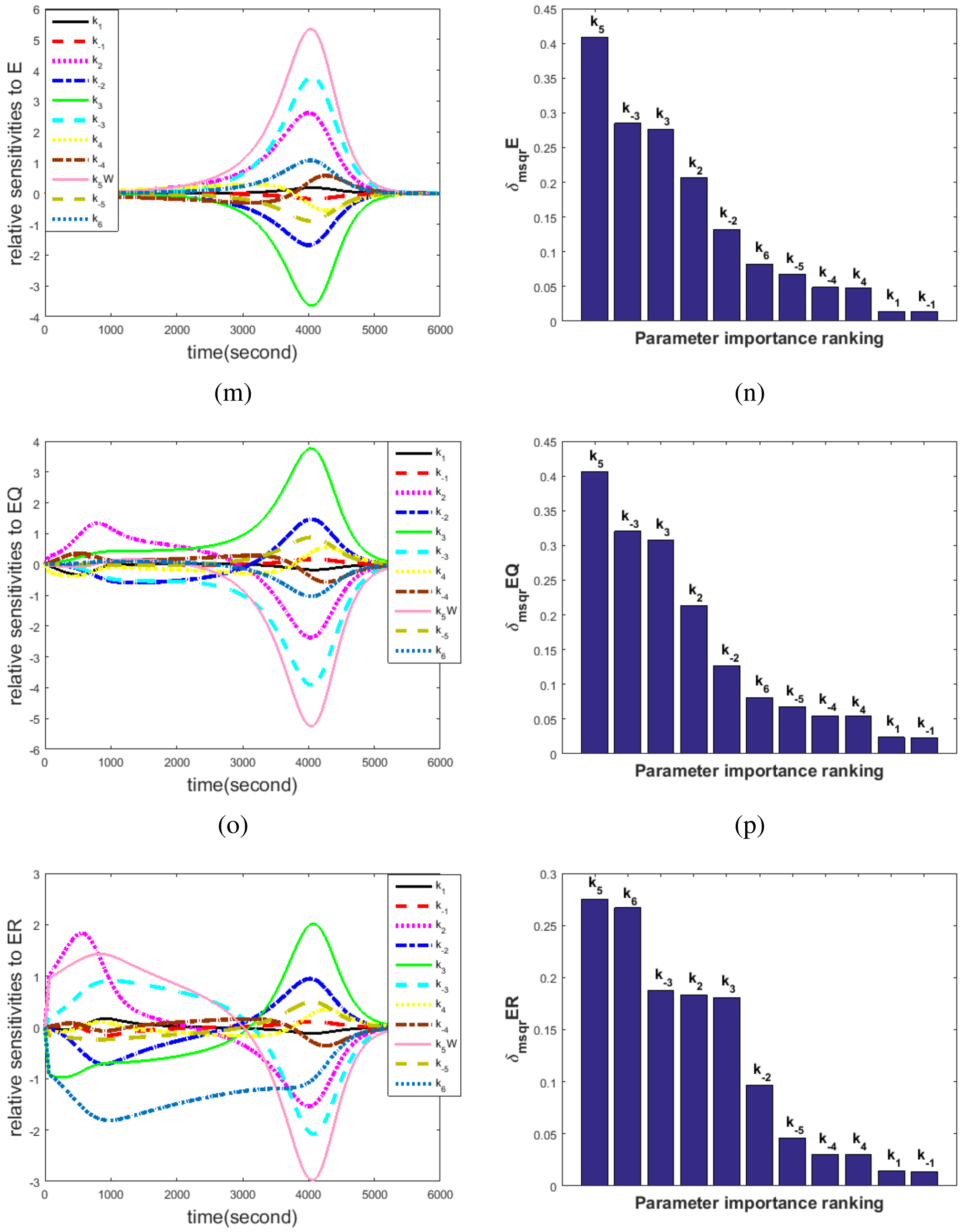
(j)


relative sensitivities to E*
time(second)
3
2
1
0
-1
-2
-3
-4
0
1000
2000
3000
4000
5000
6000
k_1
k_{-1}
k_2
k_{-2}
k_3
k_{-3}
k_4
k_{-4}
k_5W
k_{-5}
k_6

(k)


$\delta_{msqr}E^*$
Parameter importance ranking
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
k_5
k_{-3}
k_2
k_3
k_{-2}
k_6
k_{-5}
k_{-4}
k_4
k_1
k_{-1}

(l)


relative sensitivities to E
time(second)
k1
k-1
k2
k-2
k3
k-3
k4
k-4
k5W
k-5
k6
0
1000
2000
3000
4000
5000
6000
-4
-3
-2
-1
0
1
2
3
4
5
6
δmsqrE
Parameter importance ranking
k5
k-3
k3
k2
k-2
k6
k-5
k-4
k4
k1
k-1
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
relative sensitivities to EQ
time(second)
-6
-5
-4
-3
-2
-1
0
1
2
3
4
δmsqrEQ
Parameter importance ranking
relative sensitivities to ER
time(second)
-3
-2
-1
0
1
2
3
δmsqrER
Parameter importance ranking
k5
k6
k-3
k2
k3
k-2
k-5
k-4
k4
k1
k-1
0
0.05
0.1
0.15
0.2
0.25
0.3

(m)

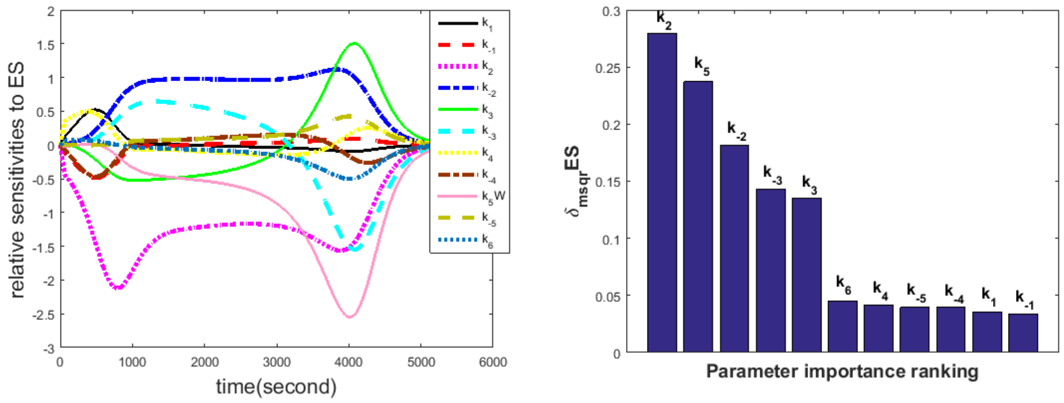
(n)

(o)

(p)

(q)

(r)



(s) (t)

Fig. 4.6 Relative sensitivities to state variables and the corresponding parameter importance ranking based on integrated effect over time horizon. Here k_5 in the right column means k_5W

k_2, k_5W, k_3, and k_{-3} are more important than other parameters. The ranking order of k_2 on state variable R is different from that on Q and N. It is also found that the relative parameter sensitivities to the enzyme complexes, of very small concentrations, are also high and their corresponding parameter importance rankings are similar to those measurable state variables, except for state variable ER where k_6 is also identified to be an influential parameter. In Figure 4.6 we can see that the magnitudes of kinetic parameter sensitivities to state variable S and Q are higher than that to P, N and R, which imply that measurements of S and Q could lead to more informative data for parameter estimation.

In addition, Table 4.4 shows the parameter pair correlations based on the calculation from FIM as follows:

$$R_{i,j} = \frac{cov\left(k_i, k_j\right)}{\sqrt{FIM_{ii} \times FIM_{jj}}} \tag{4.4}$$

It can be found that more than ten parameter pairs are highly correlated where their correlation coefficients are larger than 0.99, such as $[k_1, k_{-1}]$, $[k_3, k_{-3}]$ and $[k_5W, k_{-5}]$.

The integrated effects of parameters along time horizon on all five measurable state variables are shown in Figure 4.7, where k_2, k_5W, k_{-2} and k_{-3} are relatively more important than other parameters when we set the threshold to be 0.04 (one unit change of parameter

Table 4.4 Parameter pair correlations for the enzyme reaction system model

	k_1	k_{-1}	k_2	k_{-2}	k_3	k_{-3}	k_4	k_{-4}	k_5W	k_{-5}	k_6
k_1	1										
k_{-1}	**-0.9996**	1									
k_2	0.7966	-0.7833	1								
k_{-2}	-0.8064	0.7980	-0.9679	1							
k_3	-0.7265	0.7178	-0.9467	0.9883	1						
k_{-3}	0.7343	-0.7258	0.9337	-0.9872	**-0.9958**	1					
k_4	-0.3366	0.3332	-0.6095	0.7339	0.8141	-0.8265	1				
k_{-4}	0.4265	-0.4199	0.6976	-0.7953	-0.8617	0.8788	-0.9840	1			
k_5W	0.7551	-0.7464	0.9554	**-0.9948**	**-0.9987**	**0.9955**	-0.7893	0.8416	1		
k_{-5}	-0.7551	0.7464	-0.9554	**0.9948**	**0.9987**	**-0.9955**	0.7893	-0.8416	-1	1	
k_6	0.7591	-0.7499	0.9627	**-0.9959**	**-0.9975**	**0.9937**	-0.7755	0.8314	**0.9996**	**-0.9996**	1

would cause 20% change in state variable). In parameter identification, more efforts should be put on important parameters in order to reduce their estimation error, while a less accurate result can be accepted for those parameters which have little effect on model predictions. However, it should be noted that parameter correlations are not considered in the above relative sensitivity analysis. For example, k_3 and k_{-3} are highly correlated which can be seen in Table 4.4. Therefore, the practical parameter identifiability needs to be analysed in order to find identifiable parameters by considering both parameter influence on model outputs and correlations between parameter pairs. This will be discussed in the following section.

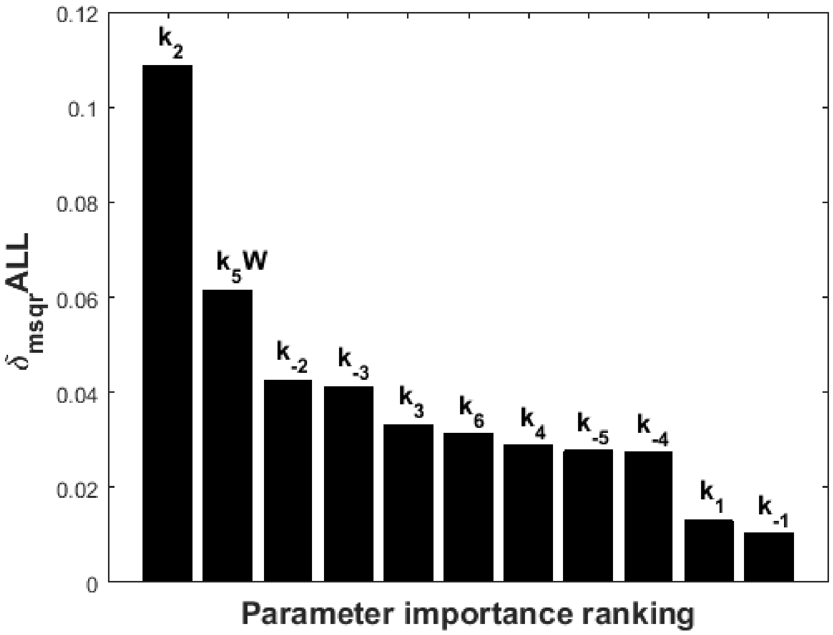


Fig. 4.7 Parameter ranking based on relative sensitivities for all state variables

4.3 Identifiability analysis for enzyme reaction system

4.3.1 Structural identifiability analysis

Identifiability analysis is a process to uncover problems with model structure and parameter identifiability. Through the identifiability analysis the minimum model structure and identifiable parameters can be determined. This work is important as it can improve both the interpretation of the model's functionality as well as the identifiability of parameter estimates. In most biochemical networks, it is usually the case that model parameters cannot be uniquely estimated due to the highly interconnected nature of the system. Identifiability analysis can be performed to check whether this problem is mainly caused by the model structure. When the model is proposed, the structure identifiability problem is concerned with the ability of determining unique parameter estimation solutions under perfect measurement condition (noise free and continuous in time and space). It is thus related to model structure and type of input stimulus, but independent of parameter values.

In this work, the generating series approach is applied to investigate the structural identifiability of the benchmark enzyme reaction system. This method is confined to models which are linear in the input stimulus, given as follows:

$$\begin{aligned} \dot{\mathbf{X}}(t) &= \mathbf{f}(\mathbf{X}(t), \boldsymbol{\theta}) + \sum_{i=1}^{n_u} g_i(\mathbf{X}(t), \boldsymbol{\theta})\, u_i(t) \\ \mathbf{Y} &= \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}) + \boldsymbol{\xi}(t) \end{aligned} \tag{4.5}$$

where g_i is the coefficient associated to the i-th input variable. For model equations (4.5), the observation can be expanded in such a way where the series coefficients are $\mathbf{h}(\mathbf{X}(t_0), \boldsymbol{\theta})$ and its Lie derivatives

$$L_{f_{j0}} \cdots L_{f_{jk}} \mathbf{h}\left(\mathbf{X}\left(t_0\right), \boldsymbol{\theta}\right) \tag{4.6}$$

where

$$L_{\mathbf{f}} \mathbf{h}\left(\mathbf{X}\left(t_0\right), \boldsymbol{\theta}\right) = \sum_{j=1}^{n} g_j\left(\mathbf{X}\left(t_0\right), \boldsymbol{\theta}\right) \cdot \frac{\partial}{\partial x_j} \mathbf{h}\left(\mathbf{X}\left(t_0\right), \boldsymbol{\theta}\right) \tag{4.7}$$

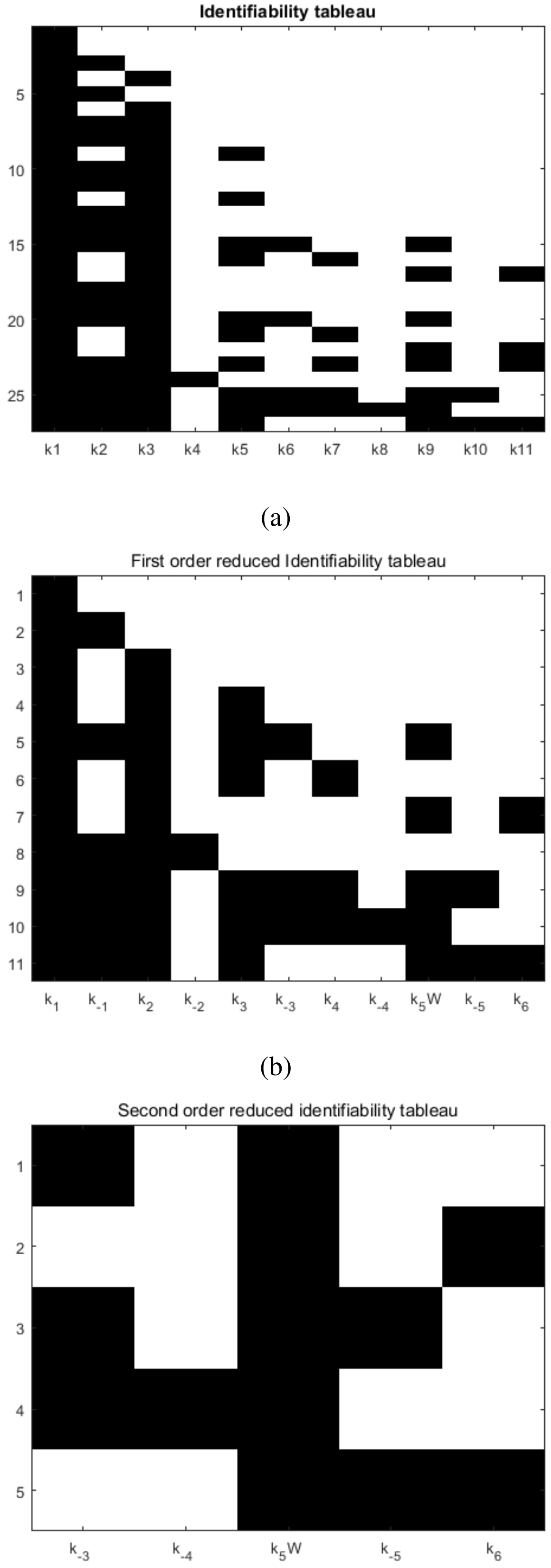
in which g_j is the j-th component of $\mathbf{g}$. Then all the series coefficients are combined in a vector and a sufficient condition assuring the model to be identifiable is that there is a unique solution of parameters θ to this series coefficient vector.

The generating series approach combined with the identifiability tableau (Balsa-Canto et al., 2010; Chiş et al., 2011) is applied to the enzyme reaction system model. Ideally it is always possible to obtain a full rank Jacobian matrix for the power series coefficients because the number of Lie derivatives is infinite. Through the numerical steps proposed in Appendix A, the reduced and the minimum tableau can be obtained. The model parameters can then be determined as globally structurally identifiable (G.S.I.), locally structurally identifiable (L.S.I.) and non-identifiable (N.I) depending on whether they have an unique solution, more than one solution or are non-solvable. The corresponding identifiability tableaus are shown in Figure 4.8, with all the five measurable state variables included. Each row represents one series coefficient determined by the Lie derivative and each column represents one model parameter. In such a tableau, each black grid denotes that the series coefficient in that row contains the non-zero element with respect to the model parameter in the corresponding column. In Figure 4.8a there are 27 non-zero series coefficients with respect to the eleven model parameters, which are obtained by the Lie derivative computations. Figure 4.8b shows a reduced tableau where eleven necessary rows are selected which can guarantee full rank of the Jacobian matrix. In this tableau a unique non-zero element in a given row means that the model parameter in the corresponding column can be identified, and this identifiable parameter can then be removed from the tableau. The elimination of a column (parameter) will lead to a reduced tableau with new unique non-zero elements. This process will continue, iteratively, until the tableau cannot be further reduced. For the enzyme reaction system, the final minimum tableau is shown in Figure 4.8c, in which the five remaining parameters need to be further checked to see whether they are structurally identifiable or not. When all the five measurable state variables are included in this observation, all of the eleven model parameters can be determined as globally structurally identifiable.

One should note the fact that it is possible not to measure unimportant state variables while more efforts will be taken on more important state variables. In this case, not all five measurable state variables will be measured during experimentation. Therefore, the structural identifiability with different combinations of measurable state variables are tested, which is shown in Table 4.5. It can be found that measurement of at least three state variables can guarantee that model parameters are at least locally identifiable. In the case that only two state variables are measured, it is found that parameters can be locally identified if state variable S is selected to be measured, which implies that the substrate in this system is an important reactant. However, due to high computational load, the structural identifiability by measuring only one state variable cannot be tested.

4.3.2 Practical identifiability analysis

Models that are structurally identifiable do not necessarily mean their parameters can be identified in a practical situation because experimental data are usually sparse and inevitably corrupted with noise, and also parameters may be highly correlated with each other. It is therefore required to analyse the practical identifiability of the model so that the identifiable parameter set could be determined. In Section 2.5.2 different approaches for analysing practical identifiability have been discussed. In this work, three different parameter subset selection methods, principal component analysis based method (Degenring et al., 2004),



Identifiability tableau

(a)

First order reduced Identifiability tableau

(b)

Second order reduced identifiability tableau

(c)

Fig. 4.8 Identifiability tableaus based on generating series approach

Table 4.5 Structural identifiability analysis with different number of measurable state variables by using generating series approach combined with identifiability tableau

Measurement variables	Lie derivative order	G.S.I.	L.S.I	N.I.
one state variable	derivative number is too high to calculate			
two state variables	derivative number is too high, S is an important measurement set			
three state variables	7	at least locally identifiable		
four state variables	6	at least k_1, k_{-1} and k_2 are globally identifiable	the remaing 8 parameters	/
S, P, N, Q, R	5	all parameters	/	/

collinearity method (Brun et al., 2001) and orthogonalisation method (Yao et al., 2003), are applied and compared in this enzyme reaction system model.

Experimental setup

In this enzyme reaction system model, there is no prior experimentation or intended experiment strategies that are implemented or designed for data collection. Therefore, it is assumed that twenty equally spaced sampling points along the whole process (seconds, the sampling time profile shown as [initial time : sampling interval : final time]) will be collected. All five measurable state variables (S, P, N, Q, R) will be sampled at the same time points. The initial input condition and the nominal parameter values in Table 4.2 and Table 4.3 will be used in this simulation. As no experimental data is available, the average value of each state variable for the intended twenty sampling points and the nominal values of kinetic parameters will be used to scale the local parameter sensitivities.

Simulation results and discussion

(a) *Orthogonalised method*

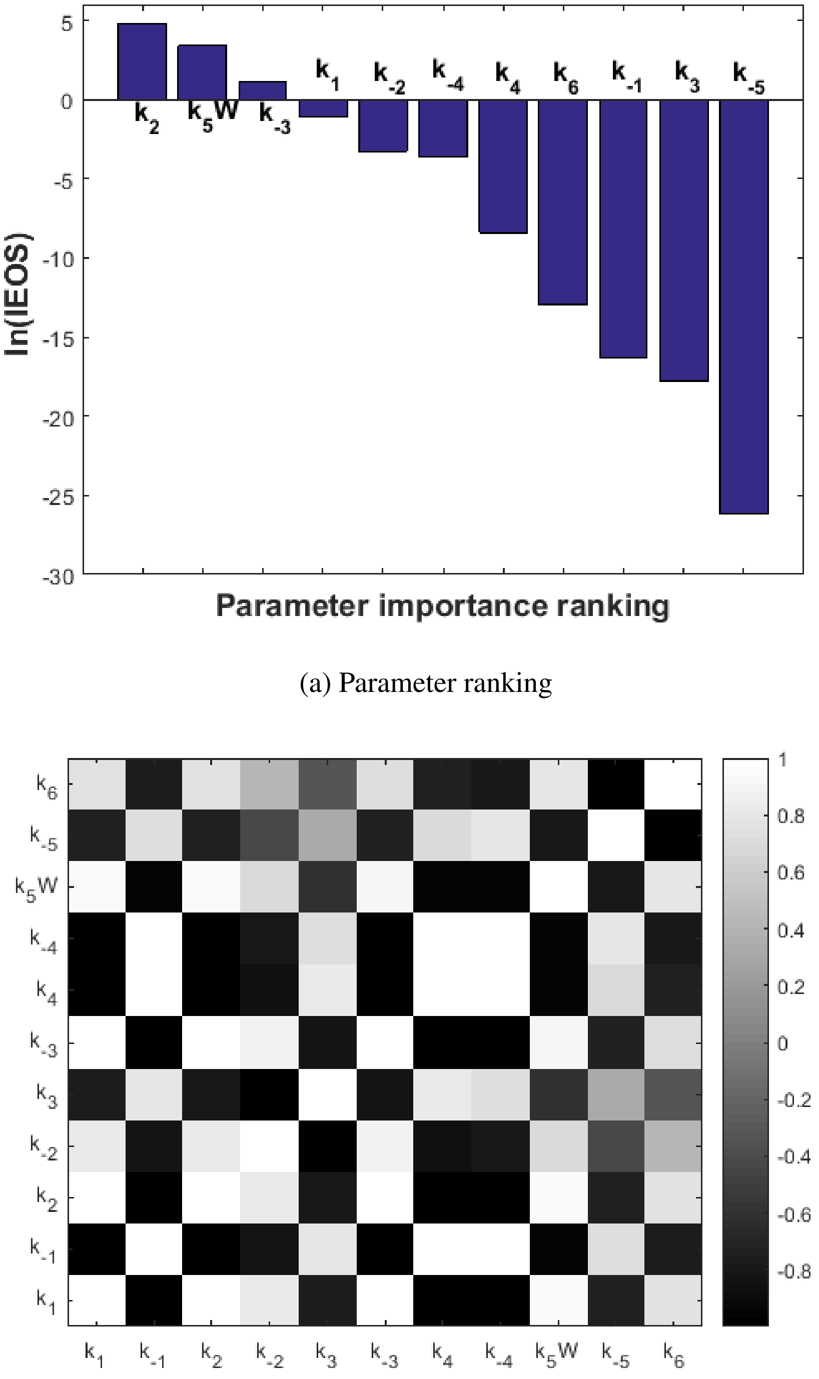
Firstly, the orthogonalised forward selection method is applied (detailed procedure is given in Appendix A) to rank the parameters based on their orthogonal sensitivities to model outputs. The ranking result and parameter correlations are shown in Figure 4.9. Here 'IEOS' represents the integrated effect of model parameters to model outputs by using orthogonalisation based local sensitivity analysis. It can be found that k_2, k_5W, k_{-3} and k_1 are the most important parameters which have higher averaged effect on model outputs. All these parameters make positive effect for the generation of R. If we set the threshold to be 0 for $\ln(IEOS)$ which is to say 100% increase (or reduction) of one parameter will have impact on all model state variables no less than 10% on average, then only three parameters, k_2, k_5W, and k_{-3} are practically identifiable.

(b) *PCA based method*

Similar to the orthogonalised forward selection method, the parameter subset selection based on PCA is also performed in a sequential process. The non-dimensional parameter sensitivities are first calculated and then the FIM is constructed based on the scaled parameter sensitivities, of which the eigenvalues and eigenvectors are obtained. Then we start from the smallest eigenvalue, the parameters are ranked from least estimable to most estimable (backward selection) by selecting the parameter with the largest weight in the corresponding eigenvector. In this way, the least estimable parameter is selected, followed by the second one and so on, until all kinetic parameters are ranked. In contrast to the backward elimination, the forward selection method starts from the largest eigenvalue of FIM and selects the most estimable parameter.

Table 4.6 Parameter ranking based on the integrated effect on model outputs with PCA

Methods	Parameter ranking (in descent order)
Backward elimination	$k_1 > k_3 > k_2 > k_{-3} > k_{-2} > k_{-4} > k_4 > k_6 > k_{-1} > k_5W > k_{-5}$
Forward selection	$k_2 > k_3 > k_{-2} > k_{-3} > k_1 > k_{-4} > k_4 > k_6 > k_{-1} > k_{-5} > k_5W$



(a) Parameter ranking

(b) Parameter pair correlations

Fig. 4.9 Parameter importance ranking by orthogonalised sensitivity analysis & parameter correlations

Both methods are applied to this enzyme reaction model and the ranking results are given in Table 4.6. It can be seen that different selection strategies of model parameters can lead to different results. Also, the ranking order of parameters from PCA analysis is very different from that generated from the orthogonalisation approach. This is because in PCA the parameter correlation is not well considered. In addition, a threshold is also required for the PCA based method to select the number of parameters which are deemed as estimable.

(C) *Collinearity index method*

Rather than the PCA and the orthogonalisation based approaches which both try to find one unique identifiable parameter subset, the collinearity index method tests all possible combinations of parameter subsets and figures out potentially possible estimable parameter subsets for estimation. Following the procedures described in Appendix A, the collinearity index of this enzyme reaction model is calculated. The threshold for the collinearity index is set to be 10 (4.10). There are in total 2036 different combinations of parameter subsets and only 123 parameter subsets fulfil that threshold where the maximum number of identifiable parameters is three (shown in Figure 4.10 and zoomed in Figure 4.11). However, this method only tests column-wise parameter correlations and the effect of each parameter on model outputs is not considered. It is possible to choose a parameter subset from which the collinearity index value is very small but have little effect on model outputs. Furthermore, the computational load could be expensive when dealing with complex biochemical system with a large number of model parameters.

The practical identifiability analysis for this enzyme reaction model indicates that under the intended experimental conditions the maximum number of identifiable parameters is three. Based on the orthogonalised forward selection method, k_2, k_5W, and k_{-3} are selected to be identifiable parameters, which will be used in the later on OED. It should be noted that all these methods are based on the assumed available experimental data or intended experimental setting. The change of experimental conditions may change identifiable parameter subset.

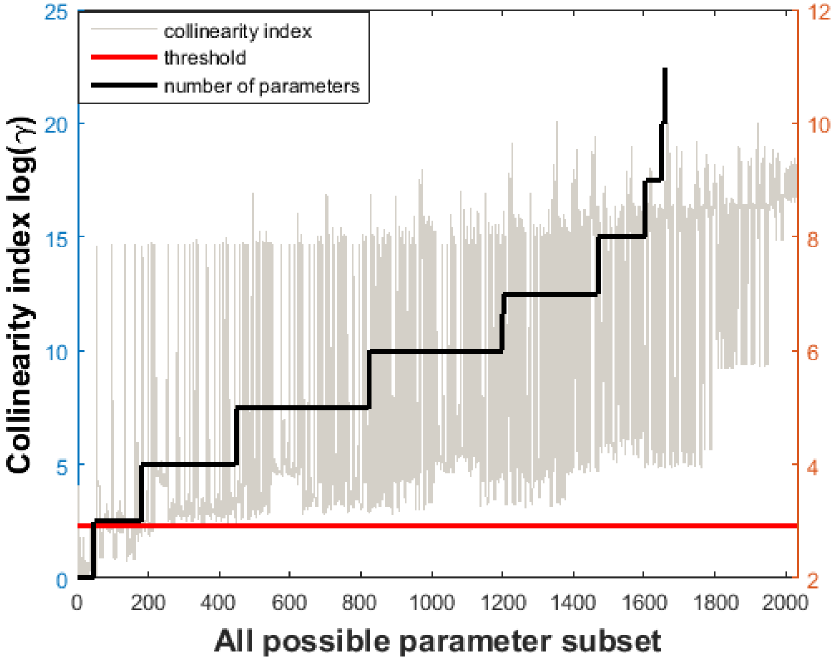


Fig. 4.10 Collinearity index for all possible parameter subsets

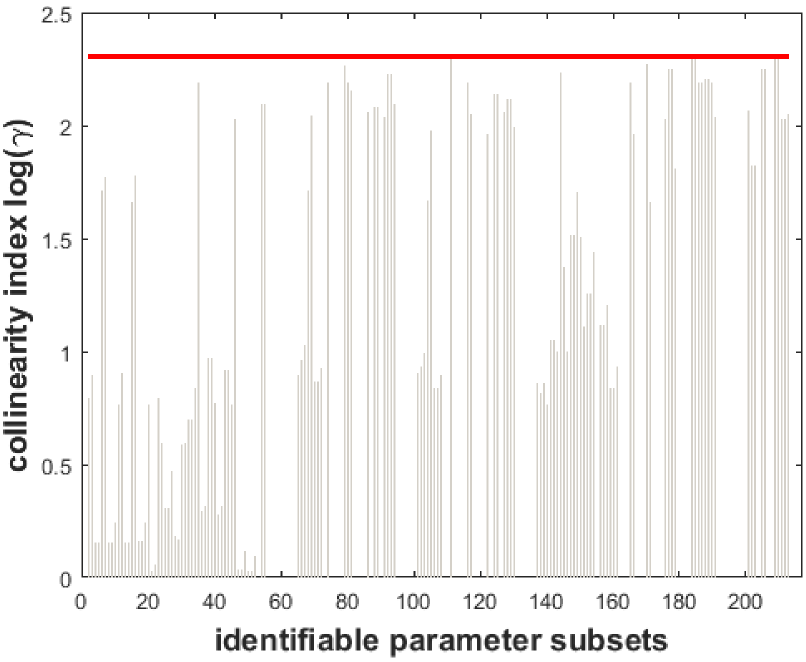


Fig. 4.11 Identifiable parameter subsets for the enzyme reaction system model

4.4 Enzymatic biodiesel production system

For enzymatic biodiesel production the conventional method is to use immobilized biocatalyst in order to improve enzyme recovery and increase stability, which allows for re-use. However, the immobilised carrier, as well as the immobilisation process significantly increases the cost of the biocatalyst, which further necessitates the re-use of enzymes for the process to be competitive. Recently the use of liquid lipase formulations for enzymatic biodiesel production has attracted more attention because it can make a significant reduction in the biocatalyst cost. Figure 4.12 briefly describes the production process of biodiesel. The main processes are the esterification to generate different forms of glyceride and the transesterification to produce acyl enzyme complex which further react with methanol to get the biodiesel.

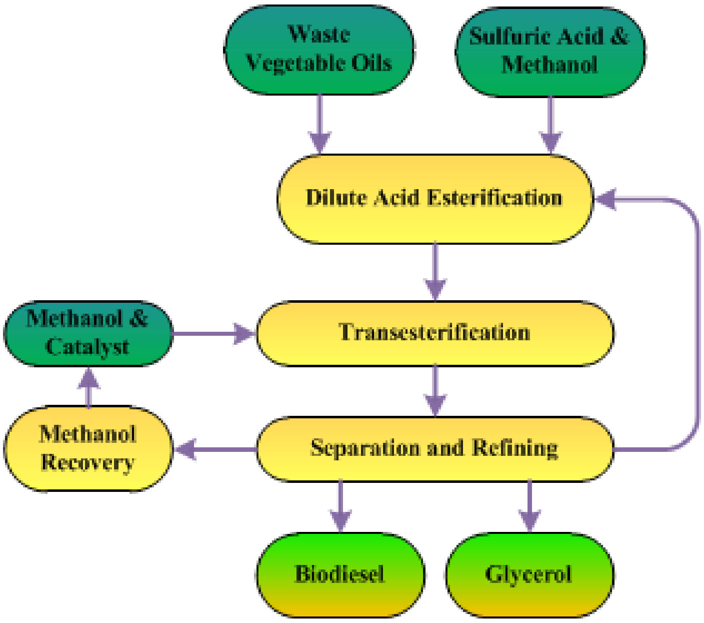


Fig. 4.12 Flowchart of the whole enzymatic biodiesel production process

Over the last few years, various kinetic models for the enzymatic transesterification of vegetable oils have been proposed. Depending on the process modelling goals, one model may offer a particular advantage over another. For engineering design, such as reactor sizing, process optimisation and control, it is desired that the kinetic model describing enzymatic biodiesel production, can predict the concentration of all the major species in the reaction. It is also essential to be able to characterise how the process responds to changes in the operating conditions over the entire course of the reaction for changes in alcohol/oil molar ratios, concentration of reactants in the reactor, enzyme loading and the area of the oil–water interface.

A kinetic model for the enzymatic transesterification of rapeseed oil with methanol using *Callera™ Trans L* (a liquid formulation of a modified Thermomyces lanuginosus lipase) was developed at DTU in Denmark (Price et al., 2014a). The model formulation is based on first principles as well as a Ping-Pong Bi-Bi mechanism. In this model the methanol inhibition, the interfacial and bulk concentrations of the enzyme are considered but not the enzyme deactivation process. The developed model is aimed at describing the effect of different oil compositions, as well as different water, enzyme and methanol concentrations; which are the relevant conditions needed for process evaluation, with respect to the industrial production of biodiesel.

While the model was developed based on theoretical knowledge and empirical procedures, those unknown parameters must be estimated from available experimental data. The use of OED methods is thus of particular importance, as these techniques allow necessary experiments to be derived in order to obtain the maximum data information, so as to improve the precision of parameter estimation.

The mathematical model describing the transesterification reaction in the biphasic oil–water system with a liquid lipase, *Callera™ Trans L*, is formulated on the basis of the following assumptions:

1. The reaction proceeds via a Ping-Pong Bi-Bi mechanism
2. The alcohol inhibition is competitive, binding to the free enzyme in place of any other reactants

3. Deactivation due to the alcohol could be ignored at low methanol concentrations

4. The interfacial and bulk concentrations of the substrate and products are the same (mass transfer from the bulk to the interface is instantaneous)

5. Non-enzymatic acyl migration is very fast so that 1- and 2-monoglycerides can be considered always stay in equilibrium

6. All reaction steps are reversible

7. All reactants are homogenous in each phase and the density of the mixture is constant

Figure 4.13 demonstrates the transesterification reaction scheme for this enzyme biodiesel production system. The free enzyme contained in the polar phase is absorbed at the water oil interface and forms the penetrated enzyme, which further reacts with triglyceride (T), diglyceride (D) and monoglyceride (M) to form enzyme substrate complexes ET, ED and EM. The enzyme substrates can be decomposed into the acyl enzyme complex and D, M and G, respectively. The acyl enzyme complex can react with water or methanol and produce the free fatty acid (FFA) and biodiesel (BD). Additionally, the competitive methanol inhibition is also considered in this reaction process.

As the interfacial and bulk concentration of the enzyme are considered during the reaction process, the free specific interfacial area needs to be calculated. The total specific interfacial area of the polar droplets can be given by:

$$a_T = \frac{6}{d_s} \cdot \frac{V_p}{V} \tag{4.8}$$

where d_s is the Sauter mean diameter of the droplets in the system (assumed to be constant, $-5.88 * 10^{-6} m$), V_p is the size of the polar volume and V is the bulk volume. Given the enzyme coverage A_e ($m^2 \cdot mol^{-1}$), the free specific inter-facial area, a_f can be calculated as:

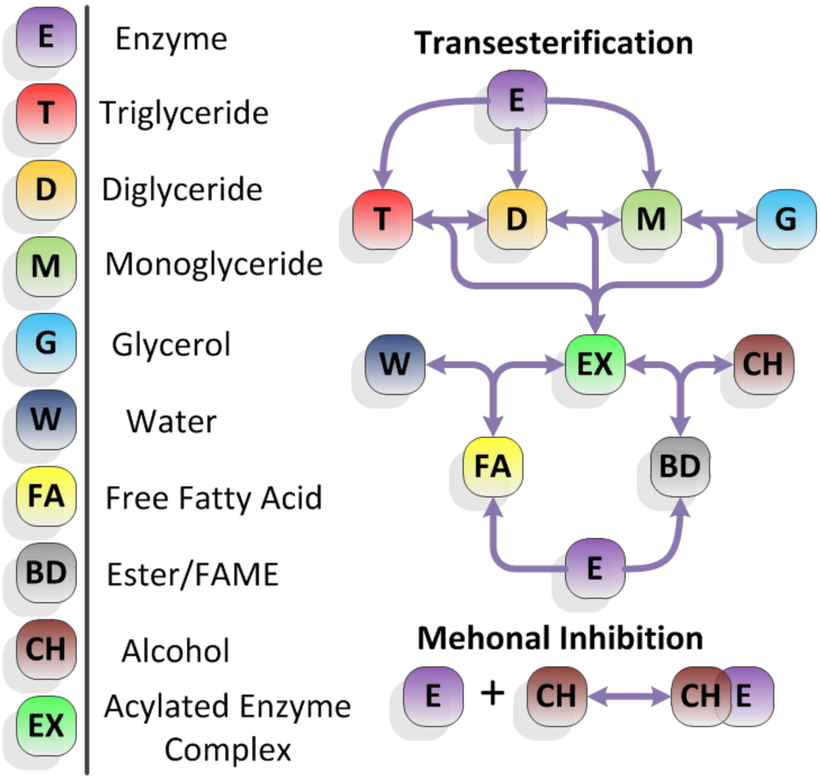


Fig. 4.13 Enzyme kinetic transesterification process

$$a_f = a_T - A_e \cdot (E + EX + ET + ED + EM + ECH) \tag{4.9}$$

Biochemical reactions and all relevant reactants during the enzymatic biodiesel production process are described in Table 4.7 and the ODEs model of this system is represented as (4.10), following the mass balance principle.

$$
\begin{aligned}
\frac{d(T\cdot V)}{dt} &= -V\cdot r_2\\
\frac{d(D\cdot V)}{dt} &= V\cdot(r_3-r_4)\\
\frac{d(M\cdot V)}{dt} &= V\cdot(r_5-r_6)\\
\frac{d(BD\cdot V)}{dt} &= V\cdot r_9\\
\frac{d(FFA\cdot V)}{dt} &= V\cdot r_8\\
\frac{d(G\cdot V)}{dt} &= V\cdot r_7\\
\frac{d(W\cdot V)}{dt} &= -V\cdot r_8\\
\frac{d(CH\cdot V)}{dt} &= -V\cdot(r_9+r_{10})\\
\frac{d(E\cdot V)}{dt} &= V\cdot(r_1-r_2-r_4-r_6+r_8+r_9-r_{10})\\
\frac{d(EX\cdot V)}{dt} &= V\cdot(r_3+r_5+r_7-r_8-r_9)\\
\frac{d(ET\cdot V)}{dt} &= V\cdot(r_2-r_3)\\
\frac{d(ED\cdot V)}{dt} &= V\cdot(r_4-r_5)\\
\frac{d(EM\cdot V)}{dt} &= V\cdot(r_6-r_7)\\
\frac{d(ECH\cdot V)}{dt} &= V\cdot r_{10}\\
\frac{d(E_{bulk}\cdot V)}{dt} &= -V\cdot r_1\\
\frac{dV_p}{dt} &= R_G+R_W, \qquad \frac{dV}{dt}=F_a
\end{aligned}
\tag{4.10}
$$

where $r_i\,(i=1,2,\cdots,10)$ are the rates of the reactions which are given in Table 4.7, and F_a, R_G and R_W are volumetric flow rate of methanol, volumetric net rates of production of glycerol and water, respectively. This model is linear in the parameters and non-linear in the state variables.

Table 4.7 Kinetic mechanism for the enzyme reaction system

Reactions		Rate of reactions
$E_{bulk} + A_f \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} E$	Enzyme in bulk absorbed in the interface	$r_1 = k_1 \cdot E_{bulk} \cdot A_f - k_{-1} \cdot E$
$T + E \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} ET$	The penetrated enzyme can react with substrate to form enzyme substrate complex ET, ED, EM (Ping) and the enzyme substrate complex forms the acyl enzyme complex and releases the first product D, M, G (Pong)	$r_2 = k_2 \cdot T \cdot E - k_{-2} \cdot ET$
$ET \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EX + D$		$r_3 = k_3 \cdot ET - k_{-3} \cdot EX \cdot D$
$D + E \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} ED$		$r_4 = k_4 \cdot D \cdot E - k_{-4} \cdot ED$
$ED \underset{k_{-5}}{\overset{k_5}{\rightleftharpoons}} EX + M$		$r_5 = k_5 \cdot ED - k_{-5} \cdot EX \cdot M$
$M + E \underset{k_{-6}}{\overset{k_6}{\rightleftharpoons}} EM$		$r_6 = k_6 \cdot M \cdot E - k_{-6} \cdot EM$
$EM \underset{k_{-7}}{\overset{k_7}{\rightleftharpoons}} EM + G$		$r_7 = k_7 \cdot EM - k_{-7} \cdot EX \cdot G$
$EX + W \underset{k_{-8}}{\overset{k_8}{\rightleftharpoons}} FFA + E$	The acyl enzyme complex can then react with water or methanol and release the second product FFA and BD	$r_8 = k_8 \cdot EX \cdot W - k_{-8} \cdot FFA \cdot E$
$EX + CH \underset{k_{-9}}{\overset{k_9}{\rightleftharpoons}} BD + E$		$r_9 = k_9 \cdot EX \cdot CH - k_{-9} \cdot BD \cdot E$
$CH + E \underset{k_{-10}}{\overset{k_{10}}{\rightleftharpoons}} ECH$	Reversible competitive methanol inhibition	$r_{10} = k_{10} \cdot CH \cdot E - k_{-10} \cdot ECH$

Note: T, D, M, G, CH, BD, W, FFA, Af, E_{bulk}, E and EX are Triglyceride, Diglyceride, Monoglyceride, Glycerol, Alcohol, Biodiesel, Water, Free fatty acid, Free interfacial area, Free enzyme bulk concentration, penetrated enzyme and enzyme complex ET represents the enzyme triglyceride complex and extends to the other complexes. Units for the concentrations are $mol \cdot L^{-1}$.

In order to evaluate values of model kinetic parameters, twelve experiments (see Table 4.8) have been conducted where the water and enzyme content were varied from 3 to 7 and 0.1 to 0.5 wt. % oil respectively. In all the experiments 1.5 equivalents (Eq.) of methanol is reacted with the rapeseed oil. One equivalent corresponds to the stoichiometric amount of alcohol need to convert all fatty acid residues in the oil to biodiesel (i.e. 1 *mol* oil : 3 *mol* alcohol). The reaction is carried out in a 0.25 *L* glass reactor with a tank diameter of 55 *mm* (T) and 2 baffles, each $0.18 \times T$ wide. The reactor is immersed in a water bath with temperature control maintained at 35°*C*. A Rushton turbine (impeller diameter 0.44 T), spinning at 1400 *rpm* provided the mixing. Initially 0.2 Eq. methanol is charged with the oil in the reactor. When the reaction mixture reaches the reaction temperature, the amount of water and enzyme to be used in the experiment, is then added to the reactor and methanol feeding started. 50 μL samples are taken from the reactor and mixed with 500 μL solvent A (acetic acid and n-heptane 4:1,000 v/v – mobile phase). Samples are then centrifuged at 14,500 *rpm* for 5 minutes and 10 μL of the supernatant is mixed with 990 μL of solvent for the HPLC analysis. 40 μL of the prepared sample is injected in the HPLC for analysis of five measurable state variables, T, D, M, FFA and BD.

Table 4.8 Preliminary experiments for data fitting and validation

Exp.	Methanol feed rate [Eq./h]	Initial dose methanol [Eq.]	Water [wt.% oil]	Enzyme [wt.% oil]
1	0.06	0.2	3	0.1
2	0.06	0.2	3	0.2
3	0.06	0.2	3	0.3
4	0.06	0.2	5	0.2
5	0.06	0.2	5	0.5
6	0.1	0.2	5	0.3
7	0.185 first 2hrs. 0.06 thereafter	0.2	5	0.2
8	0.185 first 2hrs. 0.06 thereafter	0.2	5	0.5
9	0.185 first 2hrs. 0.06 thereafter	0.2	5	0.3
10	0.06	0.4	5	0.5
11	0.06	0.4	7	0.2
12	0.06	0.4	7	0.3

The twenty unknown kinetic constants have been estimated by fitting the model equations to full time course data, using experiments 1 to 8 which covered the span of all operating conditions. The quality of the fitting has also been judged by the measurement data from experiments 9 to 12. As our research mainly focuses on the function of model-based OED to model identification, the values of those kinetic parameters estimated in (Price et al., 2014a) will be regarded as nominal parameter values, which will be used in the OED in order to make more accurate parameter estimation. The nominal values of these kinetic parameters are shown in Table 4.9.

Table 4.9 Nominal parameter values estimated from measurement data following experiments 1 to 8

Kinetic parameters	nominal values	units
k_1	4.95e4	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-1}	6.6	min^{-1}
k_2	1.69e6	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-2}	1.11e4	min^{-1}
k_3	2.07e7	min^{-1}
k_{-3}	2.20e7	$L \cdot mol^{-1} \cdot min^{-1}$
k_4	3.41e6	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-4}	1.33e7	min^{-1}
k_5	1.55e7	min^{-1}
k_{-5}	1.81e5	$L \cdot mol^{-1} \cdot min^{-1}$
k_6	9.13e4	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-6}	5.43e5	min^{-1}
k_7	7.06e6	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-7}	4.93	$L \cdot mol^{-1} \cdot min^{-1}$
k_8	2.36e4	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-8}	3.51e6	$L \cdot mol^{-1} \cdot min^{-1}$
k_9	2.54e4	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-9}	2.05e5	$L \cdot mol^{-1} \cdot min^{-1}$
k_{10}	3.23e-2	$L \cdot mol^{-1} \cdot min^{-1}$
k_{-10}	4.39e-4	min^{-1}

4.5 Preliminary analysis to biodiesel production system

In this section, the basic analysis will be conducted to this enzymatic biodiesel production system. We will look at how measurable reactants will change during the biochemical processes and how those kinetic parameters will affect the reactions. In order to facilitate further experimental design work, important model parameters that have significant effect on model outputs will be figured out from LSA based method. From the preliminary analysis, it is expected to provide a clear picture of how this mathematical model will represent the biodiesel production process and the relationship between kinetic parameters and state variables.

4.5.1 Local sensitivity analysis

Figure 4.14 illustrates the concentration time profiles of the five measurable state variables where the red star points represent the real experimental values and the blue lines describe the simulated concentration trajectories. It can be seen that the model predicts well the trends of the experimental data except for FFA which shows a clear deviation. This mismatch of FFA may be due to processes that are not taken into account. For example, the viscosity of the reaction media changes one order of magnitude over the 24 hours. Hence the parameter estimates in Table 4.9 are average values of the rate constants over the entire course of the reaction. The local sensitivity analysis is conducted in order to find relationship between model parameters and their effect on model outputs of interest.

Figure 4.15 provides an example of the relative sensitivities of model parameters to the state variable *BD*, calculated from equation (2.16). It can be seen that the influences of different model parameters on the measurable outputs are different which implies that some parameters might not be identifiable due to their little effect on model outputs. The parameter ranking results by considering various measurable states based on the LSA are

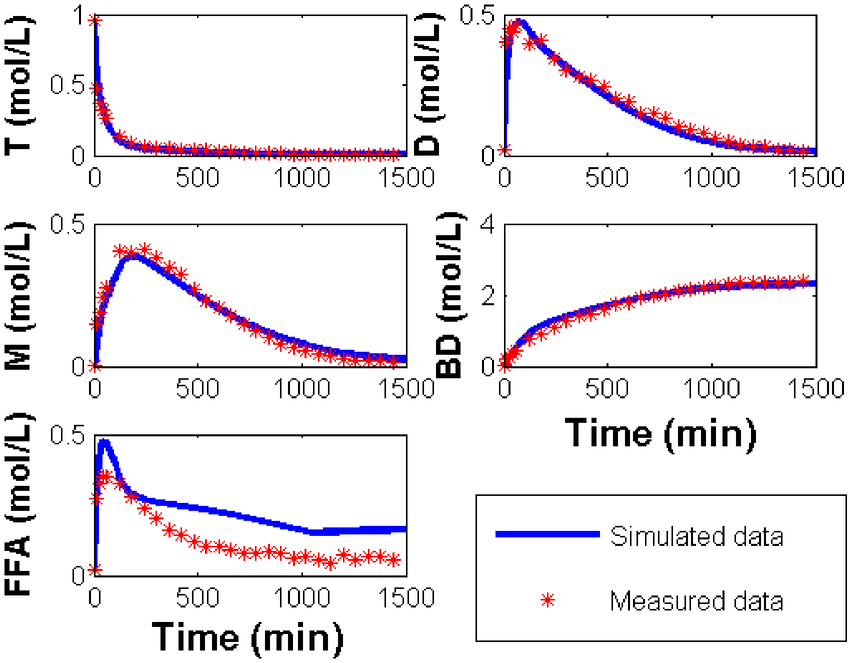


Fig. 4.14 Time profile of 5 measurable state variables in enzymatic biodiesel production system

given in Figure 4.16. It is obvious that k_6, k_9 and k_{-9} are three most important parameters for state variables T, D, M, BD and the magnitude of the integrated parameter effect on BD is very small. For state variable FFA, k_8, k_{-8} and k_9 are the most important parameters. However, the LSA does not consider correlations between parameter pairs, which may affect the estimation results.

Based on equation (4.4), the parameter pair correlations can be determined and the correlation coefficients are given in Table 4.10. It can be found that the correlations between several parameter pairs are higher than 0.99, such as parameter pairs (k_3, k_{-3}), (k_4, k_{-4}), and (k_5, k_{-5}). When we look at the relative parametric sensitivities to the output state variables FFA and T, shown in Figure 4.17, we can see that the relative sensitivity profiles of those highly correlated parameter pairs are either overlapped or symmetrical with the horizontal line corresponding to zero sensitivity level. This indicates that unique estimates of these

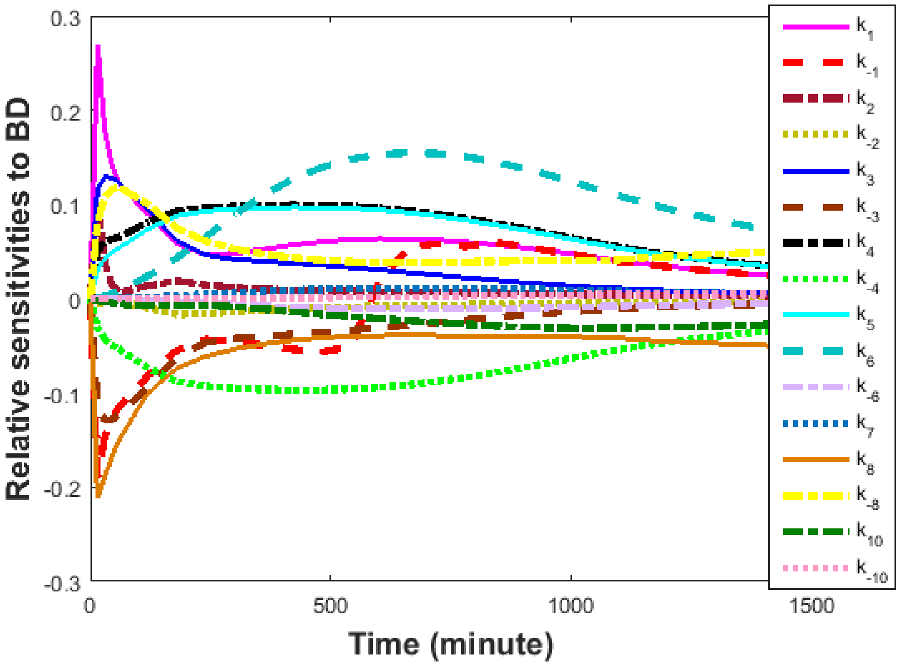


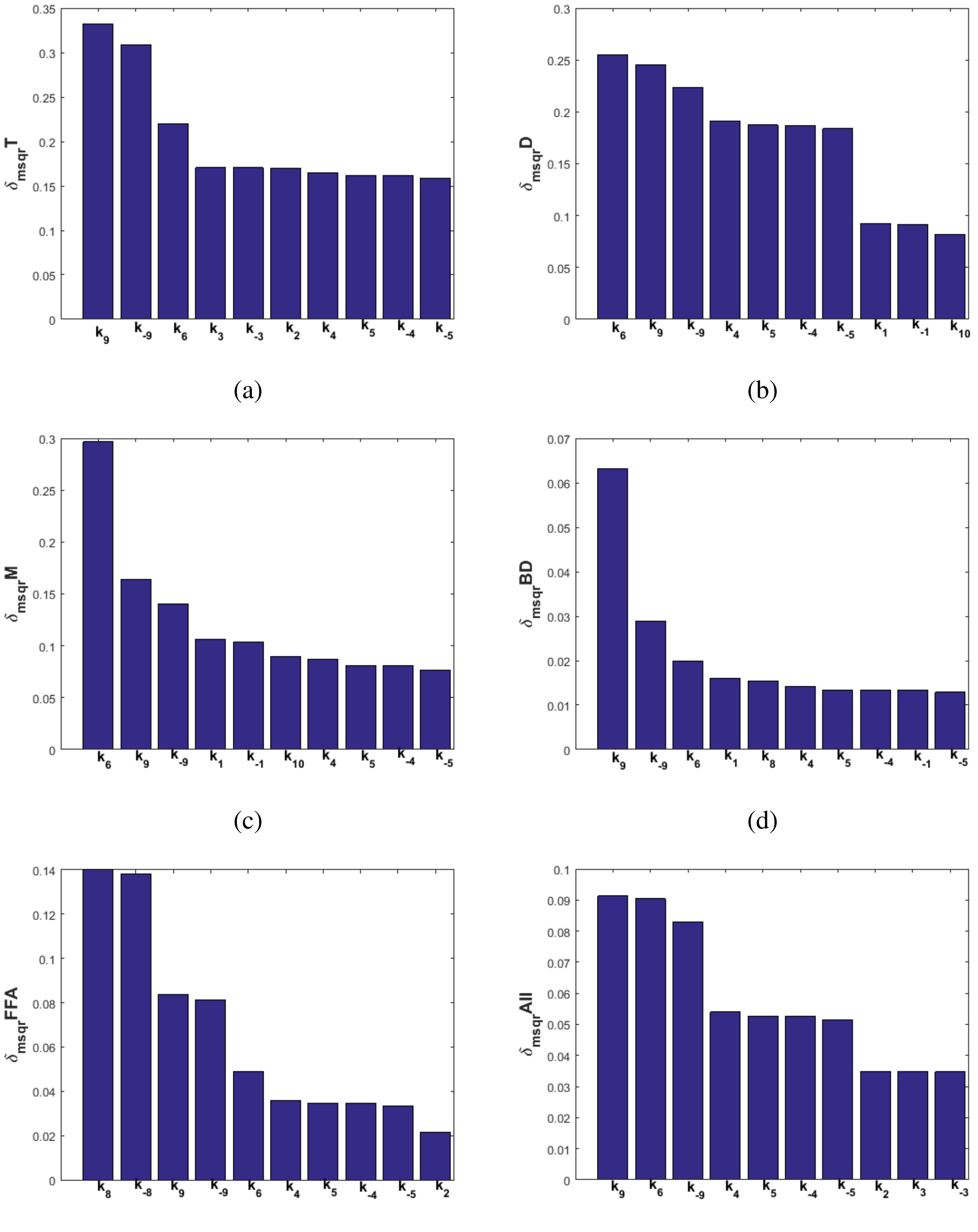
Fig. 4.15 Relative parametric sensitivities to state variable *BD* in enzymatic biodiesel production system

parameter pairs are quite difficult to obtain. The change of one parameter can be compensated by the change of another correlated parameter.

4.5.2 Parameter identifiability analysis

To further examine correlations between parameters, the collinearity index is calculated to determine estimable parameters for the system. It has been found in Figure 4.18 that the maximum number of parameters that can be estimated is ten when the threshold of the index value is set to be ten (Brun et al., 2001).

To check the result calculated from the collinearity index method, the orthogonalization based method is also applied to examine parameter correlations and to rank parameters so as to select the set of identifiable parameters. The parameter subset selection result is shown in Figure 4.19, where 'IEOS' denotes the integrated effect of the orthogonalised



(a) (b)

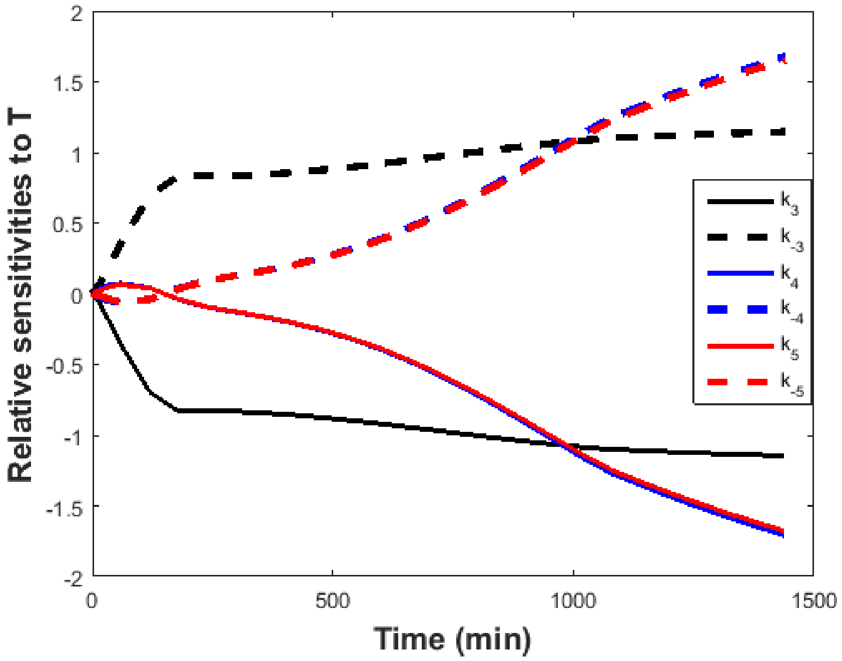
(c) (d)

(e) (f)

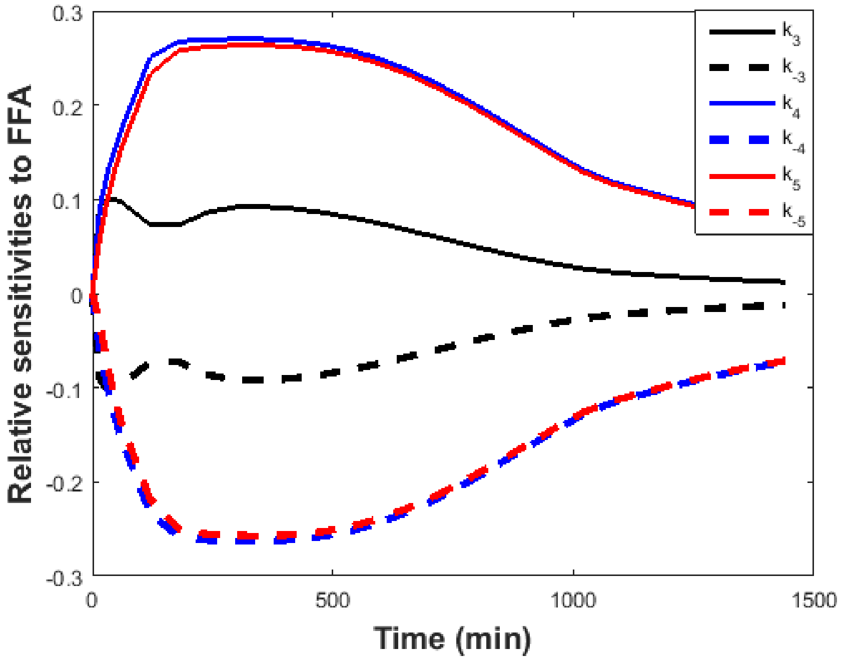
Fig. 4.16 Parameter rankings based on LSA for 5 measurable state variables in the biodiesel production system model

Table 4.10 Parameter pair correlations for the enzymatic biodiesel production system

	k_1	k_{-1}	k_2	k_{-2}	k_3	k_{-3}	k_4	k_{-4}	k_5	k_{-5}	k_6	k_{-6}	k_7	k_{-7}	k_8	k_{-8}	k_9	k_{-9}	k_{10}	k_{-10}
k_1	1																			
k_{-1}	-0.9937	1																		
k_2	0.9970	-0.9826	1																	
k_{-2}	-0.9971	0.9828	-1	1																
k_3	-0.6686	0.7477	-0.6120	0.6124	1															
k_{-3}	0.6585	-0.7386	0.6012	-0.6017	-0.9999	1														
k_4	-0.6584	0.7385	-0.6011	0.6015	0.9999	-1	1													
k_{-4}	0.6484	-0.7295	0.5904	-0.5909	-0.9996	0.9999	-0.9999	1												
k_5	-0.3206	0.2181	-0.3934	0.3928	-0.4756	0.4874	-0.4876	0.4991	1											
k_{-5}	0.3207	-0.2182	0.3935	-0.3929	0.4755	-0.4873	0.4875	-0.4990	-1	1										
k_6	-0.2354	0.3429	-0.1788	0.1794	0.8335	-0.8387	0.8388	-0.8438	-0.7019	0.7018	1									
k_{-6}	0.3009	-0.4056	0.2459	-0.2465	-0.8606	0.8650	-0.8650	0.8692	0.6587	-0.6586	-0.9976	1								
k_7	-0.2143	0.3070	-0.1359	0.1364	0.8161	-0.8225	0.8226	-0.8289	-0.8413	0.8412	0.7464	-0.7375	1							
k_{-7}	-0.3369	0.2593	-0.4090	0.4085	-0.3467	0.3578	-0.3579	0.3688	0.9256	-0.9256	-0.4331	0.3898	-0.8236	1						
k_8	-0.3215	0.2190	-0.3942	0.3937	-0.4748	0.4866	-0.4867	0.4983	1	-1	-0.7013	0.6581	-0.8408	0.9257	1					
k_{-8}	0.4742	-0.3933	0.5423	-0.5418	0.2549	-0.2672	0.2673	-0.2794	-0.9475	0.9476	0.4400	-0.3886	0.7520	-0.9840	-0.9477	1				
k_9	-0.1504	0.2245	-0.0891	0.0895	0.6339	-0.6392	0.6393	-0.6444	-0.6686	0.6686	0.5953	-0.5871	0.7798	-0.6465	-0.6682	0.5952	1			
k_{-9}	-0.1410	0.2003	-0.0935	0.0938	0.5168	-0.5207	0.5208	-0.5245	-0.5158	0.5157	0.4890	-0.4842	0.6093	-0.4854	-0.5155	0.4458	0.9513	1		
k_{10}	-0.0651	0.1764	0.0087	-0.0081	0.7855	-0.7938	0.7939	-0.8019	-0.9081	0.9080	0.9161	-0.8979	0.9233	-0.7551	-0.9077	0.7440	0.7302	0.5792	1	
k_{-10}	-0.0243	0.0742	0.0148	-0.0146	0.3673	-0.3715	0.3715	-0.3756	-0.4580	0.4580	0.3849	-0.3748	0.4858	-0.4338	-0.4578	0.4138	0.3082	0.0821	0.4752	1



(a) Selected parameter sensitivities to *T*



(b) Selected parameter sensitivities to *FFA*

Fig. 4.17 Relative sensitivities to measurable state variables T and FFA

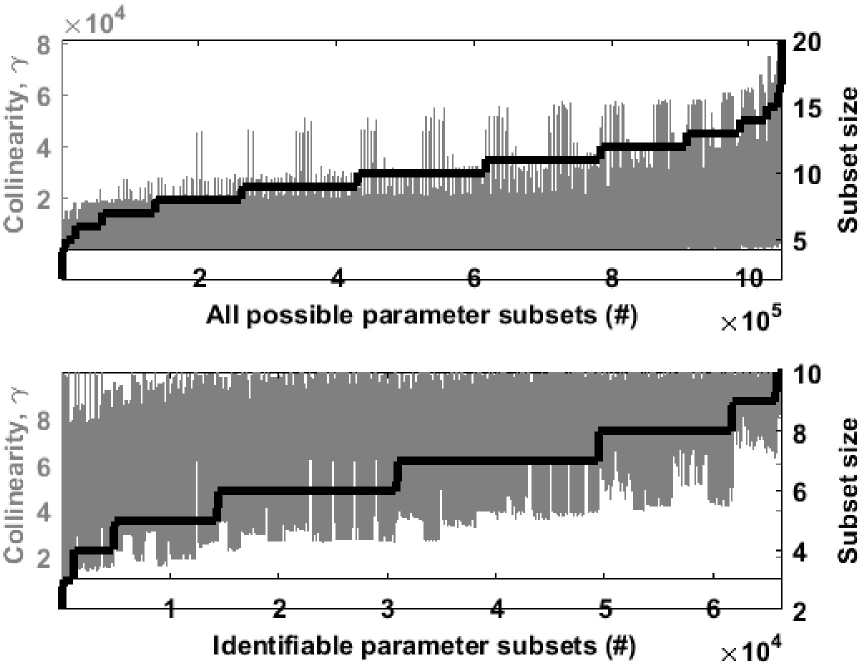


Fig. 4.18 The collinearity index values with different parameter subsets for the enzymatic biodiesel production system

parameter sensitivities to the output state variables. When we set the threshold to be 0 for $\ln(IEOS)$ (Yao et al., 2003), it can be found that those parameters above the zero line are deemed as identifiable parameters. The other parameters are kept at fixed values. Clearly both the orthogonalisation based method and the collinearity index method lead to consistent results regarding the number of estimable parameters. However, it should be noted that the collinearity index method provide several different identifiable parameter subsets but it gives no clue which parameter subset is the best one that should be estimated. Although the sequential parameter selection process based on the orthogonalised sensitivities only provide one identifiable parameter subset and the optimality of the solution cannot be guaranteed, the choice based on this method is usually good and well acknowledged in previous work (Chu and Hahn, 2012; Eghtesadi and McAuley, 2014; Yao et al., 2003). Therefore, in this enzymatic biodiesel production system we will choose these ten parameters selected from the orthogonalisation method as identifiable parameters.

Comparing the orthogonalized parameter subset selection method to the LSA based parameter ranking, which is shown in Table 4.11, it can be seen that the ranking results for parameter importance are quite different. Although k_6 and k_9 are the two most important parameters in both methods, the other parameters are ranked differently based on their importance to the output variables. This is mainly because the parameter correlations are not considered in the LSA based parameter ranking method. Therefore, the ten selected parameters, shown in Figure 4.19 should be the focus in later on OED. As the main purpose of this PhD work is to demonstrate the function of the proposed OED methods and provide guidance on the OED application to real biochemical model identification problems, only the top four most important parameters, k_9, k_6, k_8 and k_2, will be selected in further OED analysis in order to simplify the numerical calculation. This will not change the fundamental characteristics of the OED problem.

Table 4.11 Parameter importance ranking orders for the enzymatic biodiesel production system by using LSA and orthogonalised LSA

Parameter selection methods	Parameter importance ranking (descent)
LSA	k_9, k_6, k_{-9}, k_4, k_5, k_{-4}, k_{-5}, k_2, k_3
Orthogonal LSA selection	k_9, k_6, k_8, k_2, k_{-9}, k_{-1}, k_{-5}, k_{-2}, k_{10}, k_{-8}

4.6 Summary

In this chapter, two biochemical models and their preliminary analysis have been described in detail.

An enzyme reaction model which contains 10 state variables and 11 parameters has been proposed to represent a class of typical enzyme kinetic processes. The investigation of this enzyme reaction model is of particular importance as parameter identification problems of this kind of model commonly appear in chemical and biological engineering. The ODEs model

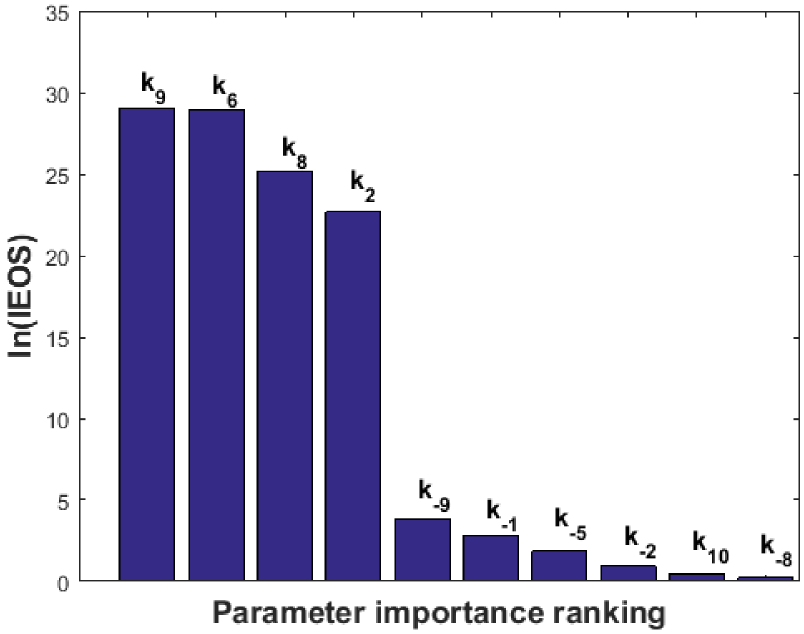


Fig. 4.19 Parameter ranking based on the orthogonal backward elimination method for the enzymatic biodiesel production system

is formulated based on mass action laws. Then preliminary analysis has been conducted on this model in order to provide a primary understanding of model dynamics. The effect of kinetic parameters on model predictions has been analysed based on the LSA. Furthermore, key parameters have been determined through the application of three different parameter subset selection methods. This enzyme reaction system model, including complete model information and typical kinetic natures, can be used as a benchmark problem for OED development.

An enzymatic biodiesel production model proposed by Price et al. has also been described in this chapter. The biochemical process has been investigated in real experimental apparatus and preliminary measurement data are available. Based on the orthogonalised sensitivity analysis, four most important parameters have been found which have the most effect on the production of biodiesel.

These two biochemical models will be used as exemplars in the following experimental design work, from which we intend to provide comprehensive experimental design procedures for the parameter estimation of biochemical processes.

Chapter 5

Single Factor Experimental Design for Biochemical Systems

This chapter investigates the development and implementation of model-based experimental design techniques to biochemical systems when a single experimental factor is considered. The FIM based experimental design methods are applied to two exemplar models which have been described in Chapter 4. The whole OED process will be described in detail through the application to two exemplar models, from which the importance of OED in system identification, particularly for parameter estimation is highlighted. We will look at how the OED problems will be formulated by considering individual design factors, i.e. input design, sampling time design and measurement set design, and how these OED problems can be solved numerically with satisfactory results. Also, different scalar design criteria will be compared and their function and limitations will be discussed. Through the case studies, guidance is provided on how to systematically apply OED methods to real biochemical processes in model identification.

5.1 OED formulations for three individual design factors

It has been mentioned in Section 3.3.2 that the model-based OED for parameter estimation is based on a scalar measure of FIM where the design factors contains initial input conditions, time-varying input variables (normally seen in fed-batch process), sampling time points, number of samplings, measurement set and others. In a particular experimental design work, the OED formulations with respect to different design factors could be quite different, and the numerical methods required to solve different OED problems are also problem dependent. In this section, we will look at how the OED problems are formulated by considering three individual design factors and how they can be solved with satisfactory results by using different numerical methods.

(1) *Input factor design*

The purpose of OED of input factors is to choose the type and duration of input stimulation/perturbations. Inputs can be fixed or time-dependent for a chemical reaction system and many other dynamic systems. When the input design factor is time-dependent, a typical option is to transfer the original OED problem into a relaxed finite dimensional non-linear programming dynamic optimisation problem by approximating the time-varying inputs with discrete form of inputs. The problem can then be solved by direct dynamic optimisation methods such as the sequential methods and the simultaneous methods.

In this work, the input factors considered for chemical reaction systems are those initial conditions of the reaction species that can be manipulated through experimental setting. The input design focuses on the design of types and quantities of the external input variables as well as the initial concentrations for reaction species in biochemical systems. Generally, the input factor design can be described as

$$\varsigma^* = \arg\min_{\varsigma \in \mathbf{U}} \psi\left(FIM(\boldsymbol{\theta}, \varsigma)^{-1}\right) \tag{5.1}$$

where ς is the vector of input design factors and $\mathbf{U}$ is the allowable input range. ψ is a scalar function of FIM, as described in Section 3.3.2. The input variables should excite the system dynamics persistently and the resulting experimental data can facilitate modelling and parameter identification. The change of input factor ς will change system dynamic responses, and the local parameter sensitivities, thus the FIM will also change. Therefore the numerical optimisation of the input design requires repeated calculation of model dynamics and sensitivities.

This input design is in general an non-convex optimisation problem that is difficult to solve to get the global solution. To obtain the optimal initial conditions of multiple inputs, in this work the particle swarm optimisation (PSO) algorithm is chosen, which has not been used in previous multi-input OED.

(2) *Optimal sampling design*

In optimal sampling design, the objective is to propose an efficient sampling strategy that selects the most informative sampling times for available measurement under the limited number of experimental runs and measurements. In most practical experimentation, typically applied sampling strategies are either the equally spaced sampling schedule or the sampling strategy with gradually increased sampling intervals as the reaction continues. Experimentalists usually choose the sampling points by heuristics or expert experience. The sampling time design has been rarely investigated. Generally the sampling design problem is to choose certain number of sampling points during an experiment and then the data information contained is measured based on the FIM. For example, the D-optimal sampling time design can be described as:

$$\begin{aligned} &\min \det \left(FIM^{-1} \left(t_1, t_2, \cdots, t_{N_{sp}} \right) \right) \\ &s.t.\ \{t_l\}_{l=1}^{N_{sp}} \in [0, T] \end{aligned} \tag{5.2}$$

N_{sp} is the total number of sampling points. This will result in an infinite dimensional non-convex dynamic optimisation problem, the solution of which requires the calculation of partial derivatives of function (5.2) with respect to time. This computation is very cumbersome and the constraints make it even more difficult to handle. Unlike the continuous sampling design method in equation (5.2), in this work the sampling time design of biochemical systems is dealt with as a discrete optimisation problem. This approach has not previously been applied for sampling time design. The available measurement points are defined and the aim is to find the best combination of a subset of data points from the whole set, which can be expressed as:

$$\begin{aligned} \boldsymbol{\zeta} &= \left\{ \begin{array}{cccc} t_1 & t_2 & \cdots & t_N \\ \omega_1 & \omega_2 & \cdots & \omega_N \end{array} \right\} \\ \boldsymbol{\zeta}^* &= \arg\min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}} \psi\left(\left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\omega_i \mathbf{S}(t_i)^T \mathbf{S}(t_i)\right)^{-1}\right) \\ & s.t. \quad \omega_i \in \{0,1\},\ \mathbf{1}^T\boldsymbol{\omega} = N_{sp} \end{aligned} \tag{5.3}$$

where $\boldsymbol{\omega} = [\omega_1, \omega_2, \cdots, \omega_N]^T$ is the weighting vector for all the available measurement points in time horizon and $\mathbf{1}$ is a column vector of ones. $N_{sp}(\leq N)$ is the total number of sampling points to be selected. ω_i can be 1 or 0. The value of 1 for ω_i implies that the corresponding time point should be selected in the measurement. It is assumed that the same sampling time profile is applied to all considered measurement variables. One should note that the predefined sampling period should be small enough so that the optimal sampling time solution is included in the predefined sampling time set.

(3) *Measurement set design*

The measurement set selection is to choose a subset of measurable state variables as measurements which can lead to the most informative data set for parameter estimation. Furthermore, it can also be used to evaluate the data information of hard-to-measure state variables via computer simulation which can help to identify potentially valuable state variables that are ignored by experimental researchers. The design of measurement set is particularly important in some circumstances where only a limited number of state variables can be measured at one time, but there are a large number of state variables in the process and it is unclear which one should be selected. For example, the experiment at a cellular level is normally conducted by measuring the quantity of proteins in the cell. However, there are hundreds or even more proteins in a cell and each time only three proteins could be measured due to the limit of the measurement technology. The measurement set design can help to decide which one should be measured so that the most useful data information can be obtained from the intended experiment. The original measurement set selection problem can be represented as (He et al., 2010; Jia and Yue, 2012)

$$\begin{aligned} & \boldsymbol{\zeta} = \left\{ \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \end{array} \right\} \\ \boldsymbol{\zeta}^* \quad = \quad & \arg\min_{\boldsymbol{\lambda}\in\boldsymbol{\Omega}} \psi \left(\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \lambda_i \mathbf{S}_i^T \mathbf{S}_i \right)^{-1} \right) \\ & s.t. \quad \lambda_i \in \{0,1\}, \; \mathbf{1}^T \boldsymbol{\lambda} = n_{sel} \end{aligned} \tag{5.4}$$

where λ_i is the weighting factor relating to the i-th state x_i which can be chosen as 0 or 1. As each state is related to one weighting term λ_i, the state measurement selection problem is then to select the number of valuable state variables for which λ_i has the value of 1. The states which have the weighting values of 1 are the selected states. The optimisation problem in equation (5.4) is an integer programming problem that can be further relaxed to an approximate, continuous optimisation problem which is given as follows (Boyd and Vandenberghe, 2004):

$$\begin{aligned} \boldsymbol{\zeta}^* &= \arg\min_{\boldsymbol{\lambda}\in\boldsymbol{\Omega}} \psi\left(\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\lambda_i\mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\right) \\ & s.t. \quad \lambda_i \succ 0,\ \forall i;\ \mathbf{1}^T\boldsymbol{\lambda} = 1 \end{aligned} \tag{5.5}$$

Here $0 \leq \lambda_i \leq 1$ is a continuous variable which means the optimal solution gives a lower bound for the original discrete optimisation problem. This vector optimisation problem can be transferred to a convex optimisation problem by applying the scalar design criterion. For instance, the E-optimal design, which minimises the largest eigenvalue of the measurement covariance matrix, can be cast as a semi definite program (SDP):

$$\begin{aligned} & min\ -t \\ & s.t. \quad \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\lambda_i\mathbf{S}_i^T\mathbf{S}_i \succ t\mathbf{I} \\ & \lambda_i \succ 0, \forall i;\ \mathbf{1}^T\boldsymbol{\lambda} = 1 \end{aligned} \tag{5.6}$$

The OED for measurement set gives an instruction of which state variables should be measured. Very little work has been reported on the design of measurement set.

5.2 Individual factor experimental design for enzyme reaction system

In this section, the application of model-based OED methods to the enzyme reaction system, described in Chapter 4 as a benchmark system, are investigated for individual factor design. The OED formulation by considering three individual design factors is discussed and numerical optimisation methods are introduced to solve the OED problems.

5.2.1 Optimal input design

Single input design

In the enzyme reaction system, the objective of the OED for the input intensity is to find the optimal initial concentrations of E_0, S_0 and N_0 that could generate the most useful data from the intended experiment. To start with, a single input variable is considered while keeping the other two at nominal values: then all three initial conditions are designed together which is referred to the multiple input design. The detailed simulation condition for all input designs is set as follows.

1. Measurable state variables: S, P, N, Q, R (all five measurable state variables will be taken for measurement in this input design);
2. Sampling time schedule: [0:300:6000] seconds (20 equally spaced measurement points for each state variable, so there are 100 measurement points in total, (20×5) in each experiment);
3. Parameters k_2, k_{-3} and k_5W are focused on the parameter estimation;
4. Measurement errors: Gaussian white noise with 10% relative error and 0.0001 $mol \cdot L^{-1}$ absolute error;

5. Input range ($mol \cdot L^{-1}$): $E_0 \in [1.5e-6\ 1.5e-4]$, $S_0 \in [0\ 1]$, $N_0 \in [0\ 1]$

(1) *Optimisation through exhaustive searching*

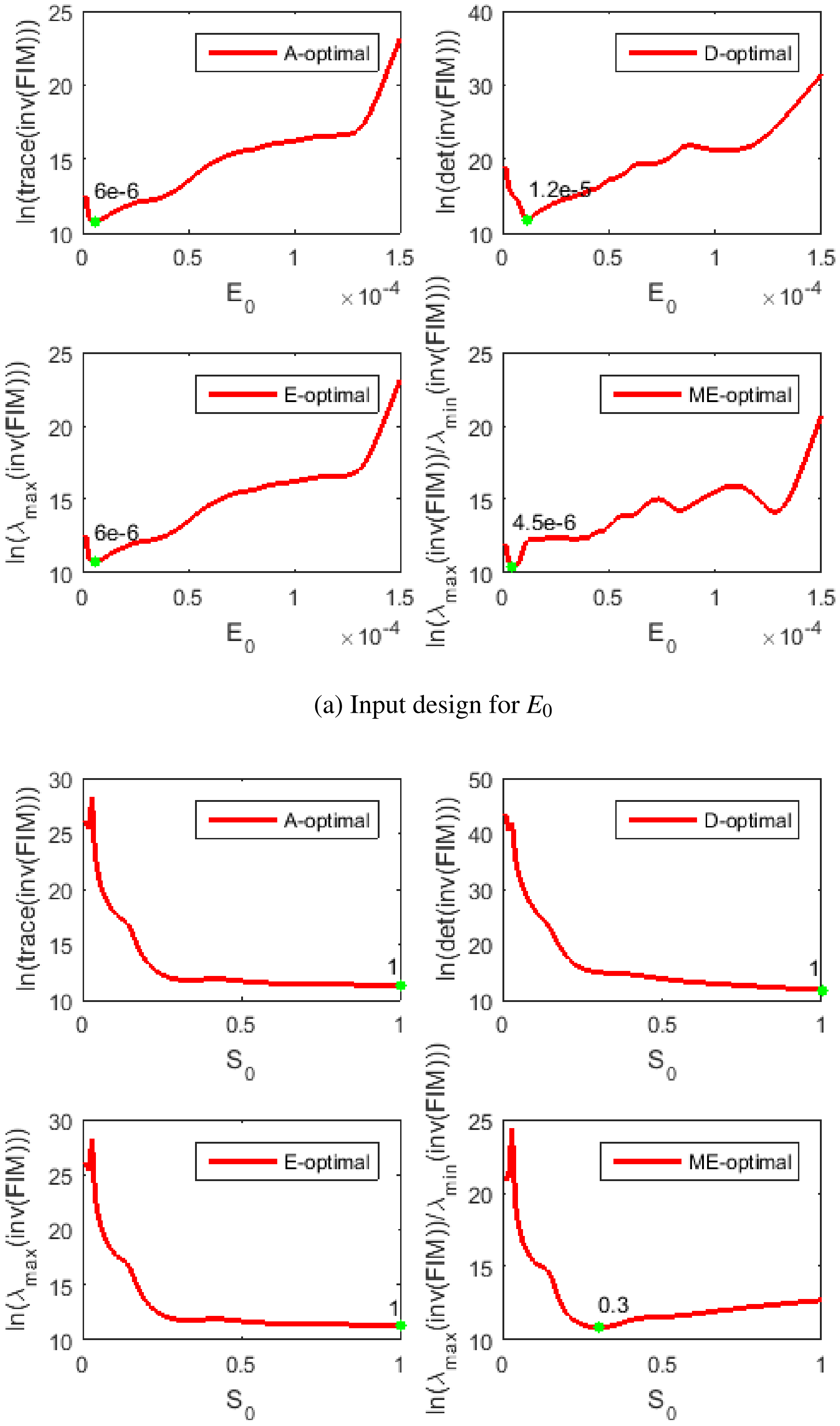
Firstly, the exhaustive search method will be used to solve the single input design problem. For each input factor of interest, it will be discretised as one hundred equally spaced values along the whole input range, which is $E_0 \in [1.5e-6 : 1.5e-6 : 1.5e-4]$, $S_0 \in [0 : 0.01 : 1]$, and $N_0 \in [0 : 0.01 : 1]$. This method can clearly show the change in OED cost function. Each time one input variable is designed to find the optimal initial concentration, the other two input variables will be kept at their nominal values. The simulation results for the three input design factors are shown in Table 5.1 and the performance index with respect to input intensity is shown in Figure 5.1.

Table 5.1 Single input design results by using exhaustive search method

Design criteria	Optimal input design results		
	E_0 ($mol \cdot L^{-1}$)	S_0 ($mol \cdot L^{-1}$)	N_0 ($mol \cdot L^{-1}$)
A-optimal	**6e-6**	**1**	**1**
D-optimal	**1.2e-5**	**1**	**1**
E-optimal	**6e-6**	**1**	**1**
ME-optimal	**4.5e-6**	**0.3**	**1**

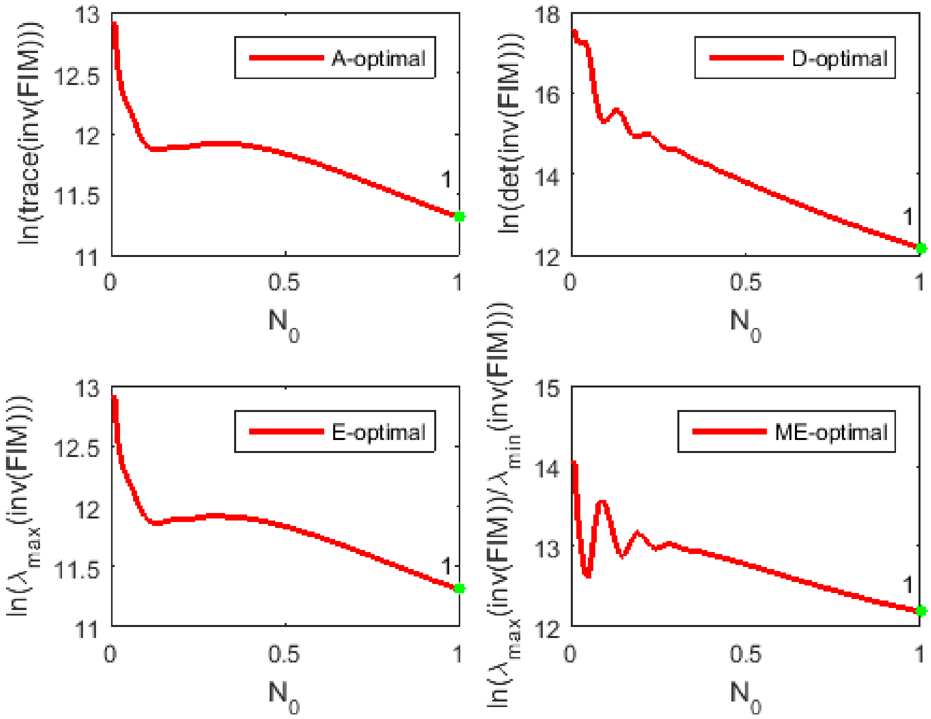
It can be seen in Table 5.1 and Figure 5.2 that all four optimal criteria for the input design of N_0 lead to consistent results, where the optimal input intensity of N_0 should be chosen as large as possible, which is 1 $mol \cdot L^{-1}$ in this case. For the design of S_0, the A-, D- and E-optimal designs suggest the value of S_0 should be set to 1 $mol \cdot L^{-1}$ while the modified E-optimal design leads to the value of 0.3 $mol \cdot L^{-1}$. In terms of the design for E_0, small input intensity is recommended.

Figure 5.1a shows that at the very beginning the increase of E_0 can make significant reduction of the OED cost function, which will then increase gradually as the increase of E_0, apart from small variations in D- and ME-optimal designs. The input design result for



(a) Input design for E_0

(b) Input design for S_0



(c) Input design for N_0

Fig. 5.1 Single input design of E_0, S_0, and N_0 with exhaustive searching method. The figures show the OED performance index w.r.t. input intensities

E_0 implies that a small amount of E_0 could catalyse the fixed amount of substrate set at the nominal condition. A large amount of catalyst will make the depletion of substrate very fast which will cause sampling data collected at the late stage of reaction to be non-informative. On the other hand, E_0 cannot be set to be too small, otherwise the reaction will be slow and make the sampling data collected at the early stage of reaction unnecessary.

Figure 5.1b and 5.1c show that the performance index of OED decreases gradually as the increase of S_0 and N_0, except for the ME-optimal design for S_0. Large values of S_0 and N_0 are suggested from experimental design. From the chemical reaction kinetic point of view, with a constant level of E_0, the value of S_0 should be set to allow moderate reaction rate in the kinetic process, neither too slow nor too fast. When S_0 is set to be 0.8 $mol \cdot L^{-1}$ and N_0 set to be 0.9 $mol \cdot L^{-1}$ at their nominal values, the optimal value of E_0 from OED is 1.2e-5 $mol \cdot L^{-1}$ in D-optimal design. When E_0 is fixed at 1.5e-5 $mol \cdot L^{-1}$ and N_0 set to be 0.9 $mol \cdot L^{-1}$, the optimal design for S_0 lead to its optimal value to be 1 $mol \cdot L^{-1}$. The OED design implies that a relatively constant ratio between E_0 and S_0 (around 1e-5) is suggested so that the enzyme is able to catalyse all substrate with moderate reaction speed, in which case informative measurement data can also be generated.

The exhaustive method, however, is computationally very cumbersome and is largely dependent on the step change of the searching variables. A small step change will require a large number of numerical calculations while a large step change might miss the global optimal solution.

(2) *Optimisation through active-set searching*

Next we will use another searching algorithm for optimal input design. The active-set quadratic programming method is applied to solve the single input design problem. The simulation condition is the same as introduced earlier. Each time only one input variable is designed and the other two input variables are fixed at their nominal values. The threshold of the minimal change of input value is set to be 1e-9 and the optimised solution is shown in Table 5.2.

Table 5.2 Single input design results for enzyme reaction system by using active-set method

Design criteria	Optimal input design results		
	E_0 design ($mol \cdot L^{-1}$)	S_0 design ($mol \cdot L^{-1}$)	N_0 design ($mol \cdot L^{-1}$)
A-optimal	**5.44e-6**	**1**	**1**
D-optimal	**1.16e-5**	**1**	**1**
E-optimal	**5.44e-6**	**1**	**1**
ME-optimal	**4.88e-6**	**0.3018**	**1**

One should note that the active-set method is a local optimisation method, the solution of which depends on the initial choice of input value. To avoid the local optimal solution, multiple initial input values should be tried in the optimisation. It can be found that both the exhaustive method (shown in Table 5.1) and the active-set method (shown in Table 5.2) lead to consistent results on the optimal values of S_0 and N_0. Also, there is little difference in the optimal values of E_0 with different design criteria from those two optimisation methods. The exhaustive method requires to define the step change of the design variable in advance and the objective function is only examined at the pre-defined set of values. Therefore the true optimal value might be missed due to the inappropriate setting of the step change. Also the convergence cannot be guaranteed and it can only provide a solution that is close to the minimum which depends on the searching step of the input. The smaller the step size, the more time is required for the searching. The active-set method which uses the gradient-based optimisation can provide a result with better precision as it can converge to the minimum solution when appropriate threshold is set (in this simulation 1e-9 $mol \cdot L^{-1}$ is given as the threshold for the input). Table 5.3 shows the function evaluation numbers of both methods in finding the optimal solution. The active-set method uses fewer objective evaluations than the exhaustive method, and it is able to generate similar optimum results as the exhaustive method.

Table 5.3 Number of function evaluations required under two numerical strategies

Design criteria	Number of function evaluations	
	Exhaustive search	Active-set method
A-optimal	100	22
D-optimal	100	4
E-optimal	100	15
ME-optimal	100	45

The single input designs with four different scalar optimal criteria have been conducted in this simulation. Following the OED results, the 95% CIs of parameter pairs are shown in Figure 5.2. In each figure, a 2D illustration is used to show the CIs for a pair of parameters. The more circular the ellipse, the less correlation is between the parameters; the smaller the area of the ellipse, the smaller is the parameter estimation error bound. From Figures 5.2a, 5.2b and 5.2c, it can be seen that the A-, E- and ME-optimal design for E_0 can make less correlated parameter estimation than the D-optimal design. By looking at Figures 5.2d, 5.2e and 5.2f, the A-, D- and E-optimal designs for S_0 can make only slightly better results in reducing parameter estimation uncertainty bound compared with the non-designed condition (nominal condition, black ellipsoid). The ME-optimal design however generates larger parameter estimation uncertainties than the non-design condition but with less correlation between parameter pairs. The design for one single input may not lead to reliable design results due to the interactions between variables.

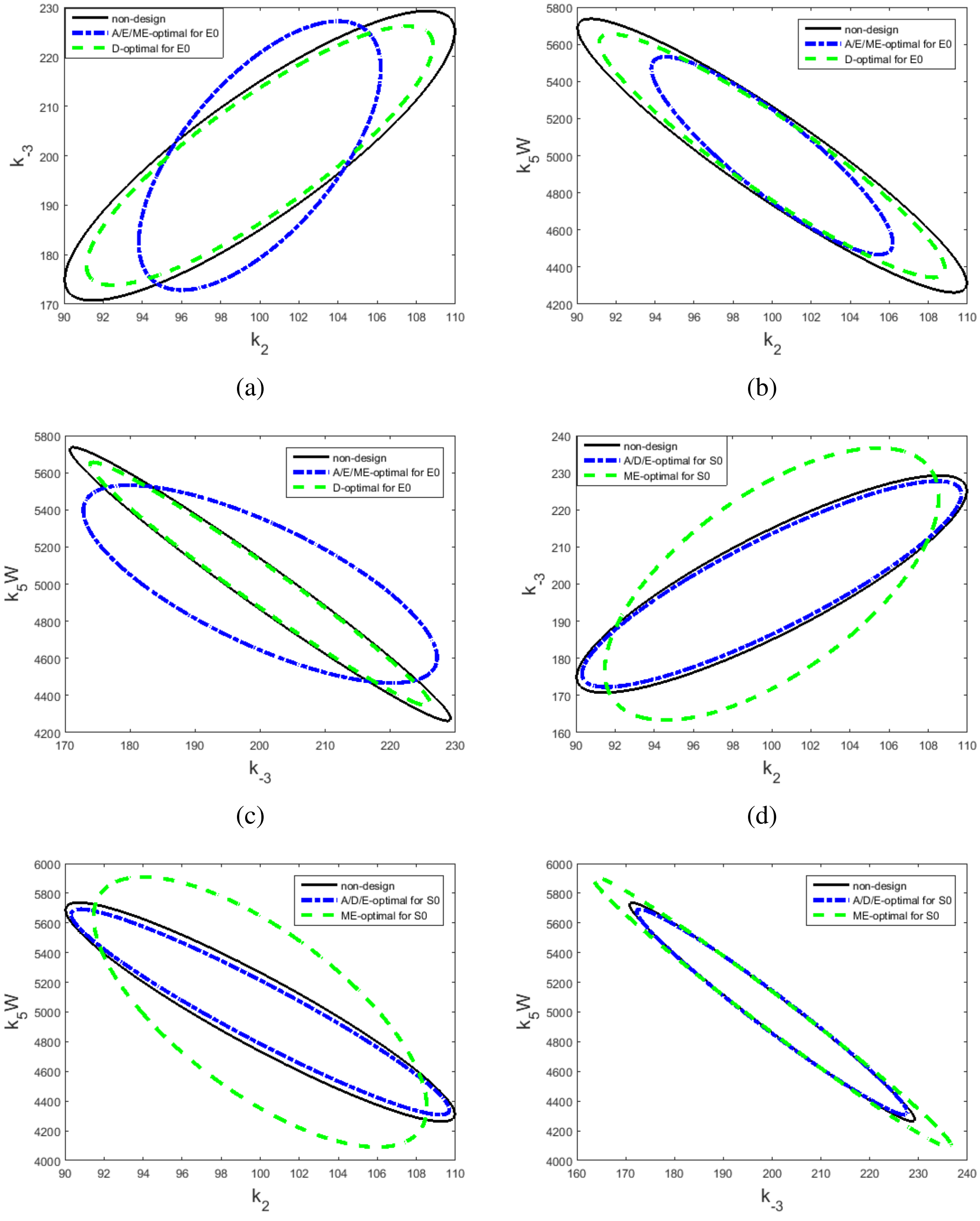
Figure 5.3 compares CIs for selected parameter pairs with different inputs. Here for each input, i.e., E_0, S_0 and N_0, the optimal solution from various design criteria is selected for comparison. That is, $E_0 = 5.88e-6$, $S_0 = 1$, $N_0 = 1$. Figure 5.3b clearly shows that the parameter estimation uncertainties of k_2 and k_5W are mostly reduced based on the E-optimal design for input variable E_0. Figure 5.3a and 5.3c demonstrate that parameter pair correlations of (k_2, k_{-3}) and (k_{-3}, k_5W) are also reduced based on the E-optimal design for E_0, which can facilitate independent parameter estimates. The uncertainty levels of model parameters from experimental design, which are calculated from equation (2.9), are given in Table 5.4. The percentage uncertainties are obtained by comparing the variations of parameters with respect to the nominal parameter values. Here the variation is calculated by half length of the CIs for each parameter. It is clearly shown that the parameter estimation precisions can be improved by using the OED, compared to the non-designed condition.

Table 5.4 Uncertainties of selected kinetic parameters from A-, E-, ME-optimal experimental design for E_0

Kinetic parameters	95% Confidence intervals		Percentage uncertainties	
	Non-design	E-opt. for E_0	Non-design	E-opt. for E_0
k_2	[92.02, 107.98]	[95.23, 104.77]	7.98%	4.77%
k_{-3}	[176.57, 223.43]	[178.70, 221.30]	11.72%	10.65%
k_5W	[4409.62, 5590.38]	[4584.41, 5415.59]	11.88%	8.31%

The ME-optimal, as described in Section 3.3.2, is used to reduce the condition number of information matrix. In other words, it is mainly focused on the reduction of parameter correlations. In this case, the parameter pair correlations in experimental conditions are shown in Figure 5.4. In the nominal condition and the condition of E-optimal design for E_0, the correlations between parameter pairs [k_2, k_5W] are very high (around 0.9), while by using ME-optimal design, the correlation between k_2 and k_5W is significantly reduced (from 0.9 to 0.7). That will help parameter estimation to obtain unique parameter values. One should note that the reduction of parameter correlations is achieved by paying a possible price for increasing parameter estimation uncertainties. These two different objectives should be balanced in experimental design. Therefore, the ME-optimal design is more suitable for the case where large correlations exist among parameter pairs.

From the single input design results, it can be seen that nominal condition is very close to the optimal designed input values for this case study system. Also, apart from the ME-optimal design which is focused on the decorrelation between parameter pairs, A-, D- and E-optimal designs lead to consistent results which can reduce the parameter estimation uncertainty

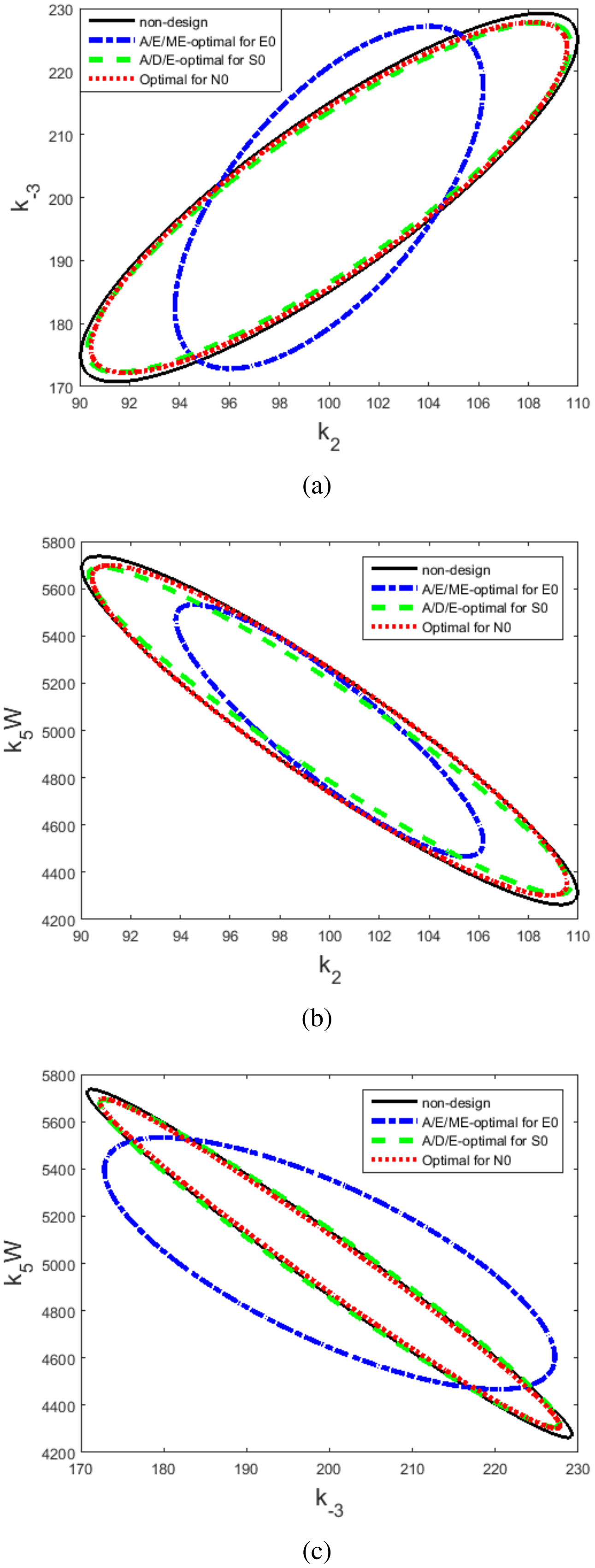


(a) (b)

(c) (d)

(e) (f)

Fig. 5.2 Comparison of different design criteria for single input design

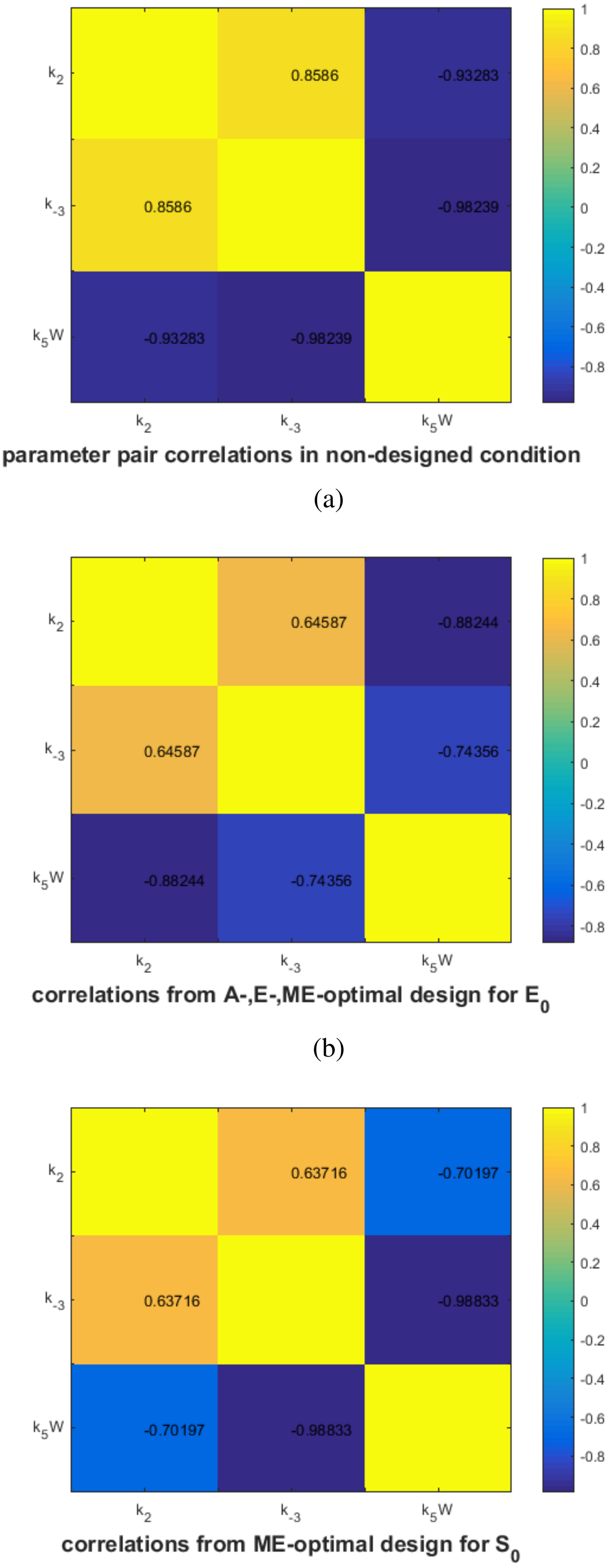


(a)

(b)

(c)

Fig. 5.3 Comparison of parameter pair CIs for different input design factors



parameter pair correlations in non-designed condition

(a)

correlations from A-,E-,ME-optimal design for E$_0$

(b)

correlations from ME-optimal design for S$_0$

(c)

Fig. 5.4 Comparison of parameter pair correlations under different experiment conditions

bound. With the single input design, the interations between the multiple input factors are not considered, which may limit the improvement of OED.

Multiple input design

The single input factor design can only provide a slightly better result than that of the non-designed condition. It can be found that the single input design problem is actually a non-convex optimisation problem. Now we look at the multiple input designs for the enzyme reaction system, where all three input factors are considered simultaneously.

The input design formulation is the same as in equation (5.1) but the dimension of design variables is increased from one to three. An efficient numerical algorithm which can lead to a solution close to the global optimum is required. In this case, two searching methods have been applied to solve this multiple input design problem. One is the active-set sequential programming method from the MATLAB optimisation toolbox, where the threshold for the input values is set to be 1e-9 $mol \cdot L^{-1}$. The other method is the differential evolutionary (DE) method (Storn and Price, 1997). DE is a population-based method that optimises the multiple input design problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. Three parameters need to be set up before the implementation of DE optimisation. The maximum number of iteration is set to be 100 for this simulation, the step size coefficient is set to be 0.9 and the crossover probability is set to be 1. The simulation results of the designed input intensities with different design criteria by using the local and global optimisation algorithms are given in Table 5.5 and 5.6, respectively.

It can be seen from Table 5.5 that the optimal input levels are largely different from each other when using different design criteria. Compared with the results from A- and E-optimal designs, the D-optimal design suggests larger values for E_0 and N_0. Also, A-optimal design suggests to use more substrate (S_0) than that from the other two designs. When using the DE

Table 5.5 Multi-input design results by using active-set method (local)

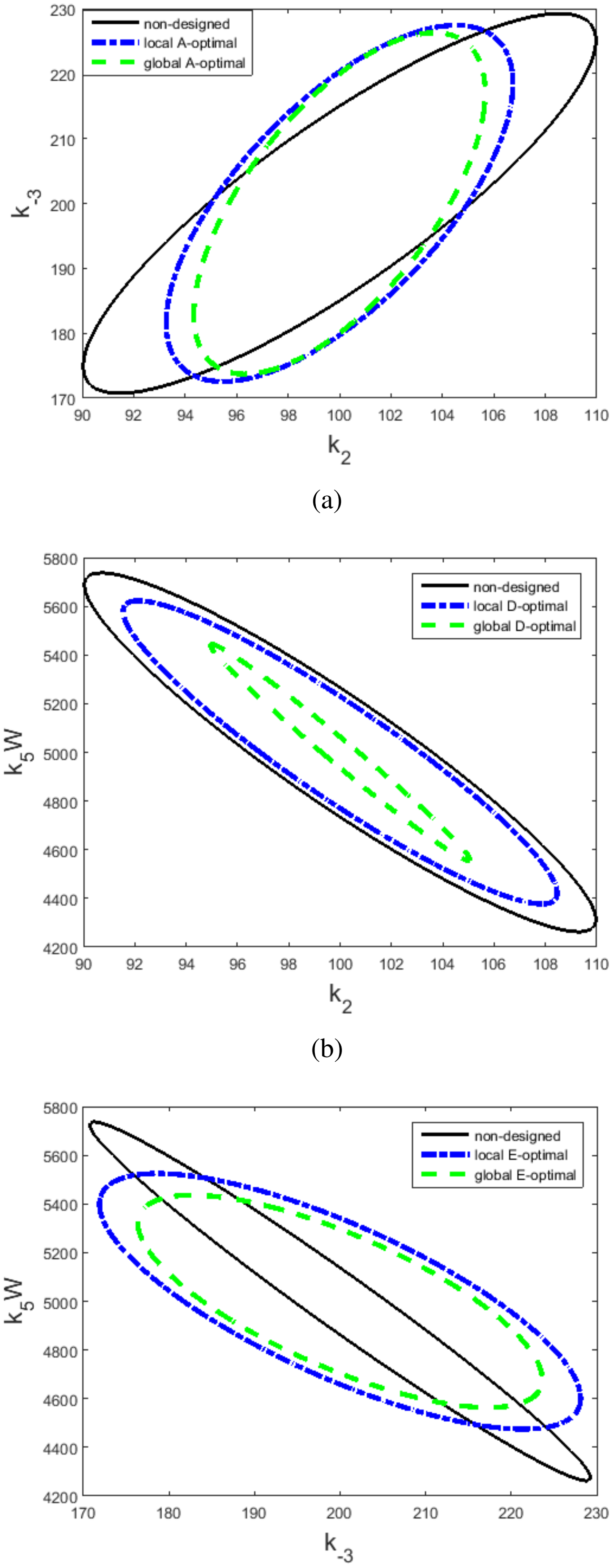
Design criteria	Design results
	$[E_0, S_0, N_0]$ $(mol \cdot L^{-1})$
A-optimal	[6.33e-06, 0.92, 0.57]
D-optimal	[1.15e-5, 0.82, 0.93]
E-optimal	[3.18e-6, 0.57, 0.49]

Table 5.6 Multi-input design results by using DE method (global)

Design criteria	Design results
	$[E_0, S_0, N_0]$ $(mol \cdot L^{-1})$
A-optimal	[4.45e-6, 1, 0.01]
D-optimal	[4.90e-6, 1, 0.07]
E-optimal	[4.27e-6, 0.94, 0.07]

global optimisation, the three input variables are calculated to be at similar levels by using the three design criteria. The local and global optimisation methods lead to very different results on initial input levels. In order to compare these two different optimisation methods, the 95% CIs for the selected parameter pairs with different design criteria are given in Figure 5.5. It is shown that the parameter estimation uncertainties in all A, D and E-optimal designs are smaller by using the DE optimisation results compared to the results from local searching. It can also be found in Figure 5.5 that both the local and global designs can lead to less correlated parameter pairs between $[k_2, k_{-3}]$ and $[k_{-3}, k_5W]$ than non-designed condition. The percentage uncertainty levels for each parameter with different experimental conditions are given in Table 5.7.

It is obvious that using the simultaneous multiple input design, the parameter estimation uncertainty bounds for k_2, k_{-3} and k_5W are all reduced more than that from single input design (see the parameter estimation uncertainty levels of key parameters in Table 5.4 and Table 5.7). The multiple input design suggests to put small amount of E_0 and N_0, but a large amount of S_0 for the intended experiment.



(a)

(b)

(c)

Fig. 5.5 Comparison of CIs with local and global optimisation methods

Table 5.7 CIs of selected kinetic parameters under different experimental conditions by multiple input design

Experimental conditions	Confidence intervals & Percentage uncertainties		
	k_2 (100)	k_{-3} (200)	k_5W (5000)
Non-designed	[92.02, 107.98] 7.98%	[176.57, 223.43] 11.72%	[4409.62, 5590.38] 11.88%
Local E-optimal	[95.60, 104.40] 4.40%	[177.49, 222.51] 11.26%	[4579.51, 5420.49] 8.41%
Global D-optimal	[95.96, 104.04] 4.04%	[181.34, 218.66] 9.33%	[4644.64, 5355.36] 7.11%

5.2.2 Optimal sampling time design

Now we consider the optimal sampling design for the enzyme reaction system. We start from a simple case that 20 distinct measurement points are required in one experiment for all five measurable state variables among 201 possible sampling points (from 0 to 6000 seconds with samplings at every 30 seconds). The initial input condition and kinetic parameters are set to be nominal values shown in Table 4.2 and Table 4.3. The *D*-optimal design criterion is applied in this case and three different numerical methods are used to search for the optimal solution. The first method is the Powell's conjugate direction method (Powell, 1964). We randomly select 20 time points from the 201 available sampling points. Then each one among the twenty will be searched for its best position while keeping the others fixed. The searching process is conducted across all the pre-selected time points and finally generate the optimal result for the sampling strategy. The second searching method is the sequential selection method. Firstly we find one single time point which can provide the most data information from the pre-defined 201 sampling time points. Then the second best time point is determined which can provide the optimal objective value combined with the first selected point. This process is iterated until all 20 sampling time points are selected. The third seraching method is the interior-point method (Byrd et al., 1999). Based on equation (5.3), all 201 pre-defined sampling time points are given small weightings that reflect their importance, where each weighting coefficient can vary from 0 to 1. Using the D-optimal design, the OED problem is a convex optimisation problem (The detailed proof can be referred to the work in (Boyd and Vandenberghe, 2004)). The interior-point method can be used to find the optimal solution with calculated weights. Then the 20 best points with the biggest weighting coefficients are selected as the optimal sampling points. The optimal results by using those three different optimisation algorithms are given in Table 5.8. Those *D*-values in the table are calculated by determinant of the inverse of FIM. The smaller the value is, the smaller uncertainty bound for the parameter estimation.

Table 5.8 Optimal sampling schedules from different numerical methods and their corresponding *D*-values. For the expression of [a:b:c], a is the starting sampling point, b is the sampling time interval and c is the final sampling point

Numerical methods	Sampling time points (seconds)	D-values
Non-designed	[300:300:6000]	4.03e-6
Powell's method	[510:30:690], [2790:30:2940], [4200:30:4380]	1.83e-5
Sequential selection	[540:30:720], 2670, [2820:30:2940], [4170:30:4380]	1.83e-5
Interior-point method	[510:30:690], [2800:30:2970], [4200:30:4380]	1.83e-5

The sampling time points is express as [starting sampling time point : sampling time interval : final sampling time point]. It can be seen that all three numerical optimisation methods lead to consistent results of the sampling time positions. The designed sampling schedule can be grouped into three parts. Several points should be selected at the early stage of the reaction because sensitivities of parameters of interest to state variables *S* and *Q* are very high during this stage. At the middle stage of the reaction (from 2,600 seconds to

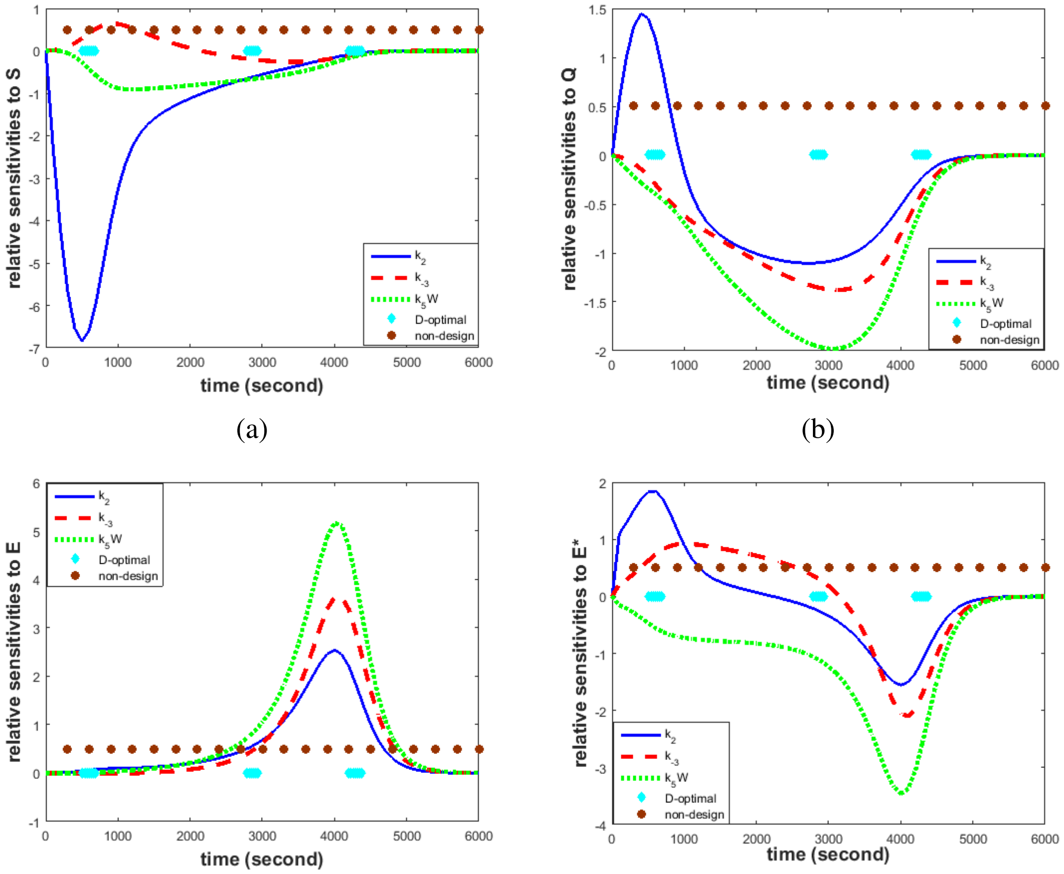
3,000 seconds) a few sampling points are suggested as the generation of product Q starts to dominate the reaction and the effect of those key parameters to state variable Q is significant. It is not obvious why sampling points should be selected at the late stage of the reaction (between 4,100 seconds and 4,400 seconds) because the relative sensitivities to S and Q are very low at that time range. However, when we look at the transient effect of key parameters on those enzyme complexes, the sensitivity values are significant between 3,500 to 4,500 seconds (shown in Figure 5.6c and 5.6d). This might be the reason that sampling points should be selected at that time range in order to provide supplementary information for these key parameters. In summary, the experimental design suggests that sampling time points should be selected during the time ranges when important parameters have significant effect on model outputs.

Figure 5.7 compares parameter estimation uncertainties under different sampling strategies. The D-optimal sampling time design by using Powell's conjugate direction method shows that proper selection of sampling time points can provide informative data that lead to effective reduction of parameter estimation uncertainty bound.

When repeated experimentation (multiple reaction runs) is allowed in the sampling schedule design, it provides more freedom for the selection of optimal sampling time points. In this case, repeated measurements at one time point may be obtained. For this enzyme reaction system, we intend to find twenty sampling points that will provide the most rich information for parameter estimation. Again the D-optimal criterion is used for OED. The simulation result is given in Table 5.9 and Figure 5.8.

Table 5.9 Optimal sampling time points and number of replicates in replicated experiments

Sampling points (seconds)	600	2910	2940	4320	D-optimal
Number of replicates	7	2	4	7	1.87e-5



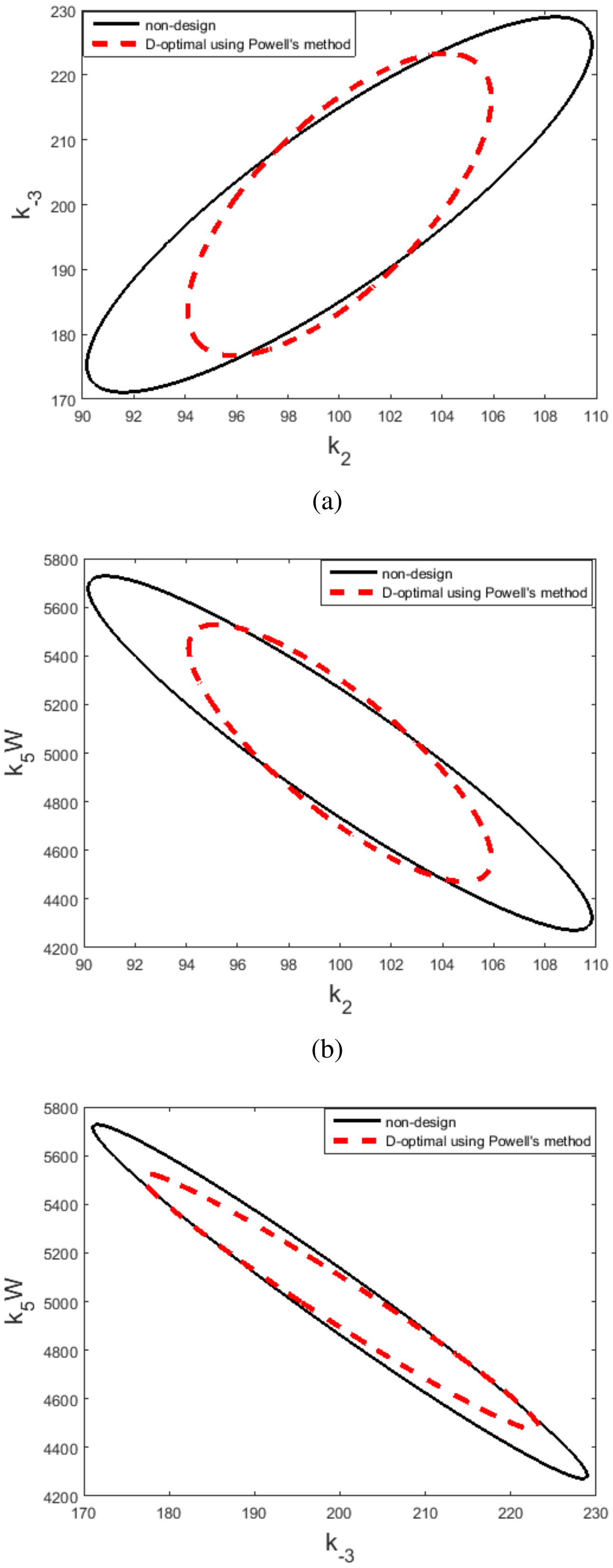
(a)

(b)

(c)

(d)

Fig. 5.6 Optimal sampling points and relative sensitivities for (a) S, (b) Q, (c) E and (d) E^*. The non-designed sampling strategy is a series of equally spaced points in brown colour. The OED sampling points are in aqua colour.



(a)

(b)

(c)

Fig. 5.7 CIs comparison of kinetic parameters under non-design sampling and the sampling schedule from *D*-optimal design

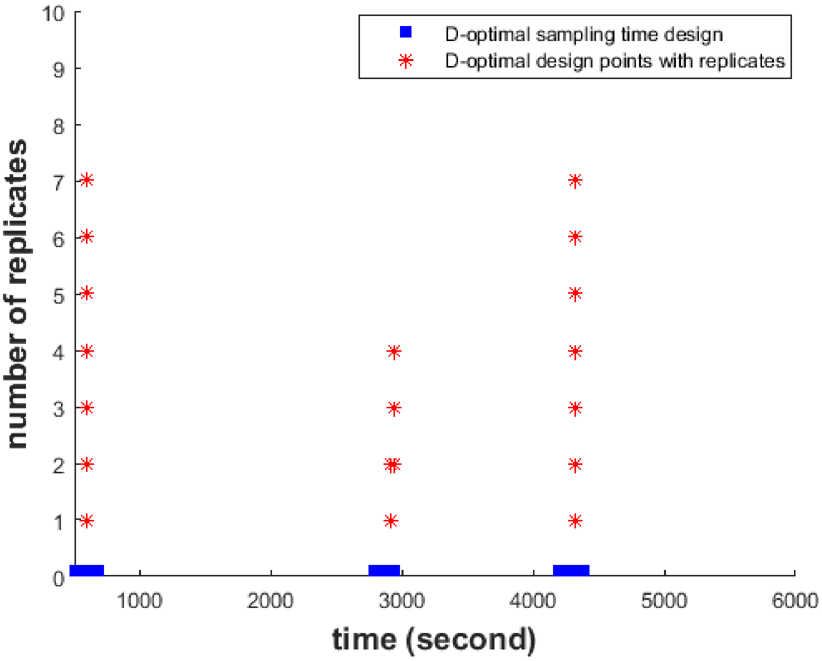


Fig. 5.8 Comparison of optimal sampling points with and without replicates. The blue dots (crowded together) are from the OED without replicates. The red dots show repeated sampling at a few time points.

It can be found that with repeated sampling allowed, only four time points need to be chosen for the measurements. They all fall in the regions of the distinct sampling time design results. These four sampling time points can be treated as the representatives at that particular time range. Because slight modifications in the suggested sampling times do not influence the performance of the scalar measure of FIM, replicated measurement at these representative points can also provide informative data for parameter estimation. From the simulation result, it can be seen that the D-value of this replicated sampling design is very close to the distinct sampling time design result. However, in practice it is not easy to make the measurement at a very precise time point. Therefore, we want to test whether these representative points can be substituted by their neighbouring time points so that the resulting sampling strategy can still provide useful data information. Within each of the three informative time regions shown in Figure 5.8, one time point is randomly selected as the measurement point with the same number of replicates taken. This process is repeated one thousand times and each time a new sampling strategy is generated. The resulted data information by D-optimal design is shown in Figure 5.9. It can be seen that most randomly generated sampling strategies can provide more informative data than the non-designed condition. One benefit of our method for sampling time design is that it can decide the informative time regions. Random selections of time points within these time ranges can generate informative data for improved parameter estimates.

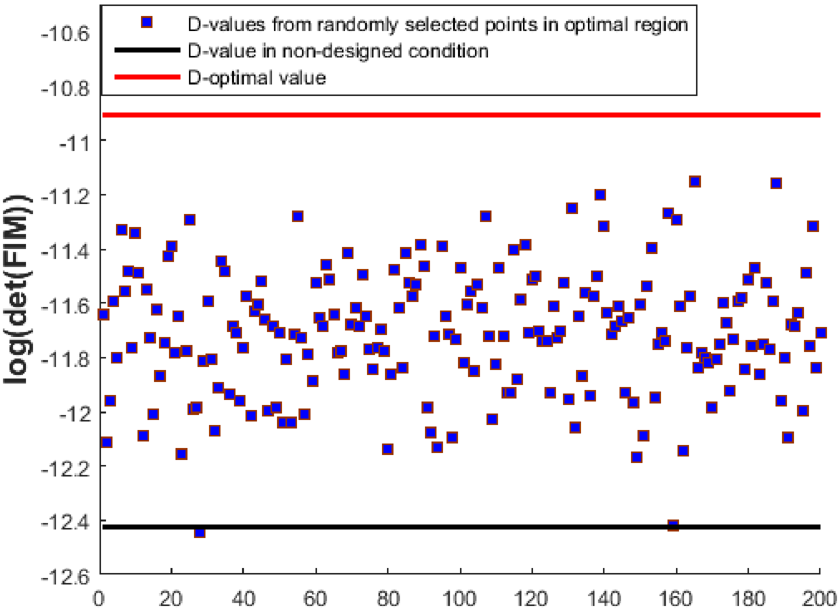


Fig. 5.9 Comparison of D-values from different sampling schedule. The figure indicates that randomly selected time points in the optimal sampling regions can normally generate more informative data than non-designed condition

5.2.3 Optimal Measurement set design

As described in Section 5.1 the measurement set design is to find out the most valuable state variables among all measurable state variables, the measurement of which would contain rich data information to facilitate parameter estimation. In this section, we will investigate how to design the measurement set in order to find the most valuable state variable as the measurement. In this enzyme reaction system, there are ten state variables in total and five of them (S, P, N, Q, R) can be measured. The simulation condition is set as follows:

1. Sampling time schedule: [0:300:6000] seconds (20 equally spaced measurement points for each state variable, so there are 100 measurement points in total, (20 × 5) in each experiment);
2. Parameters k_2, k_{-3} and k_5W are focused on the parameter estimation;
3. Measurement errors: Gaussian white noise with 10% relative error and 0.0001 $mol \cdot L^{-1}$ absolute error;

The measurement set design are conducted for two different situations, one is to consider the five measurable state variables, the other design is to consider all 10 states as possible candidates. The results are shown in Table 5.10 for the 5 states, and Table 5.11 for the 10 states. In both tables, the weighting coefficients associated with each state are given. The larger the weight is, the more significance is considered for the state.

Table 5.10 Measurement set design by considering five measurable state variables

Design criteria	S	P	N	Q	R
A-optimal	1	0	0	0	0
D-optimal	0.68	0	0	0.32	0
E-optimal	0.955	0	0	0.045	0

Table 5.11 Measurement set design by considering ten state variables

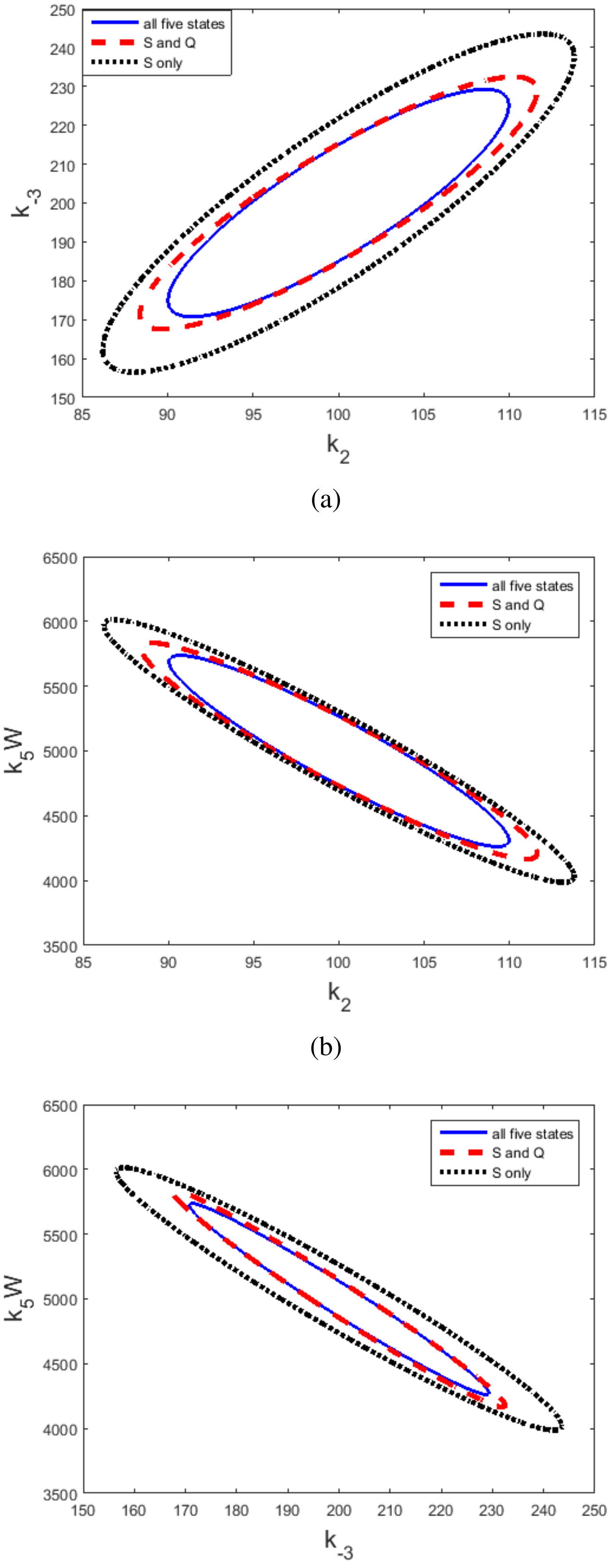
Design criteria	E, ES, E^*, EQ, ER	S	P	N	Q	R
A-optimal	0	1	0	0	0	0
D-optimal	0	0.68	0	0	0.32	0
E-optimal	0	0.955	0	0	0.045	0

It can be seen that the measurement set designs with the consideration of 5 states and 10 states can make exactly the same results on the importance of state variables. In both cases, the substrate *S* and the desired product *Q* are the two most important state variables in *E*-optimal and *D*-optimal designs. In the *A*-optimal design, only the substrate *S* is deemed as the valuable state that should be measured during experimentation. Also, it can be found that with the *E*-optimal design the substrate *S* is more important than that from the *D*-optimal by comparing the weightings. In order to compare the *A*-optimal design result with the *D*- and *E*-optimal design result, the predicted parameter estimation errors with different measurement sets are compared in Figure 5.10.

The CIs of model parameters of interest by measuring *S* and *Q* are close to that from measurement of all five state variables, whereas the measurement of *S* only can not provide imformative data. Therefore, the substrate *Q* should not be ignored in measurement.

5.3 Individual factor experimental design for enzymatic biodiesel production system

In this section, the optimal experimental design for the enzymatic biodiesel production system by considering individual design factors is investigated. Considering the competitive methanol inhibition and the requirement of the production yield of *BD*, the optimal feed of the methanol has already been defined which is given in Table 4.8. The input manipulation is therefore determined by the process requirement; only the observation needs to be designed through OED. In the following work, the optimal sampling time design and the measurement set design for this system will be analysed.



(a)

(b)

(c)

Fig. 5.10 Comparison of CIs of model parameters with different measurement sets. The data information by measuring both S and Q is close to measuring all five state variables.

5.3.1 Optimal sampling time design

The D-optimal design is employed to determine the optimal sampling strategy, i.e., at which time points to collect the measurement data. The simulation condition for the sampling time design is set as follows.

1. Measurable state variables: T, D, M, FFA, BD (all five measurable state variables will be taken for measurement in this design);

2. The minimum sampling time interval is set to be 5 minutes between two adjacent measurement. Therefore there are 301 available sampling time points in priori. 28 sampling time points is required which is the same number of sampling time points in real experimentations.

3. Parameters k_2, k_6, k_8W and k_9 are focused on the parameter estimation;

4. Measurement errors are assumed to be time independent Gaussian white noise and are equal for each observation;

The design problem can be formulated as the following optimisation problem:

$$\begin{aligned} &\max \det \left(\sum_{l=1}^{N} \mathbf{S}(t_l)^T \mathbf{S}(t_l) \right) \\ &s.t. \quad t_l - t_{l-1} \geq 5 \end{aligned} \tag{5.7}$$

Considering the 1,500 minute reaction time of the process, it is reasonable to treat these pre-defined sampling time points as representatives of their neighbouring time region because the parametric sensitivities at time points within this five minute time region are close to each other. The empirical sampling strategy in real experimentations is to take samples once every 15 minutes in the first hour and then samples once every hour in the next 24 hours. As mentioned earlier, the D-optimal sampling time design problem can be written as a convex

Table 5.12 Sampling profiles under OED and empirical sampling strategies

	Measurement time points (minutes)	D-values
D-optimal sampling	[35:5:120], [628:5:673]	3.96e-40
empirical sampling	[0:15:60], [120:60:1440]	5.39e-41

Table 5.13 Weighting coefficients for measurable state variables by measurement set selection design

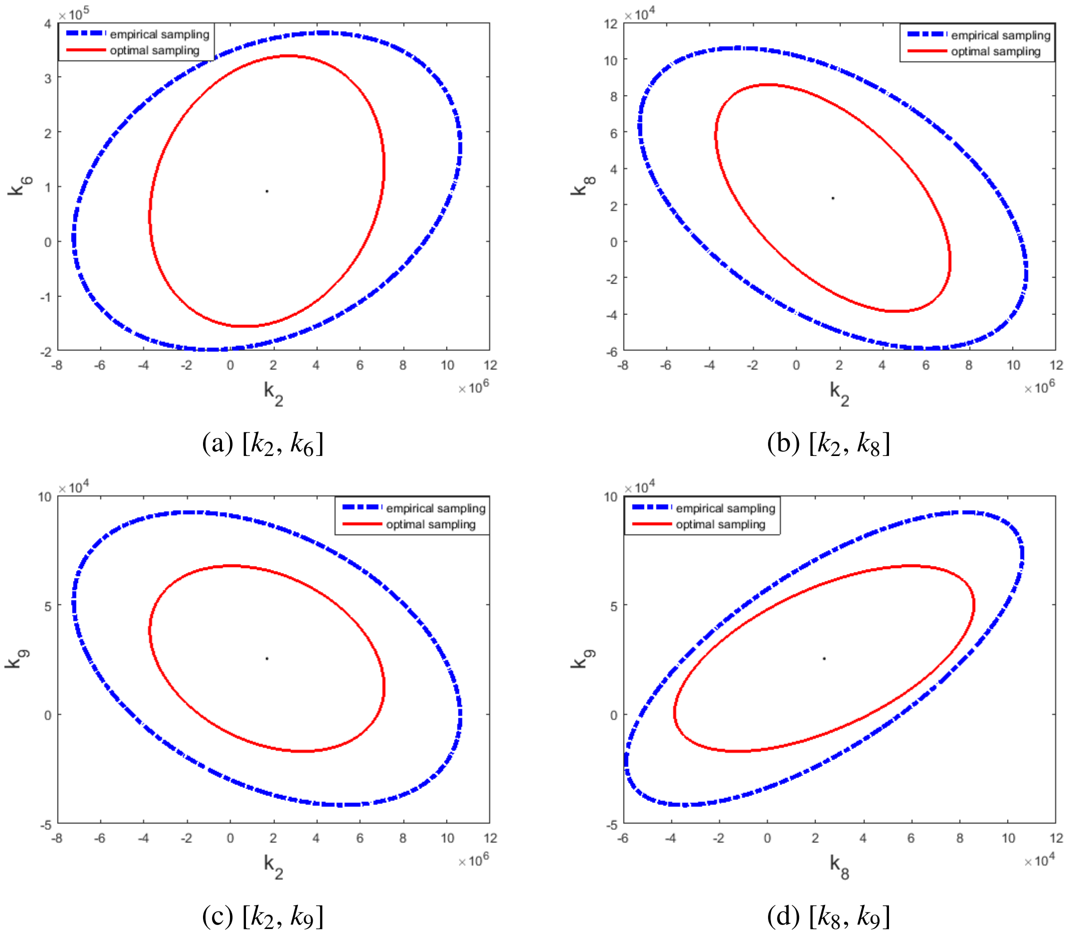
	T	D	M	BD	FFA
E-optimal	0.1675	0.1685	0.1799	0.2529	0.2312
D-optimal	0	0.1956	0	0.4688	0.3355

optimisation problem, both the Powell's conjugate direction method and the interior-point method can be used to find the global optimal solution. Within the 301 ($1500/5+1$) pre-defined time points, the optimal result is given in Table 5.12. It is suggested that the optimal sampling strategy favours those time points at the early stage of the reaction (around 35 to 120 minutes) and at the middle stage of the reaction (around 600 to 700 minutes). The CI ellipsoids for pairs of k_2, k_6, k_8, and k_9 are shown in Figure 5.11. It is not surprising that the designed sampling points lead to smaller CIs which indicate smaller bound for parameter estimation errors.

5.3.2 Optimal Measurement set design

The OED has been applied to determine the most valuable observation from the five measurable state variables. The optimal weights calculated from the E-optimal and the D-optimal designs are listed in Table 5.13.

The E-optimal design result shows that the difference of the importance among the five measurable state variables are small and all five measurable state variables should be



(a) $[k_2, k_6]$

(b) $[k_2, k_8]$

(c) $[k_2, k_9]$

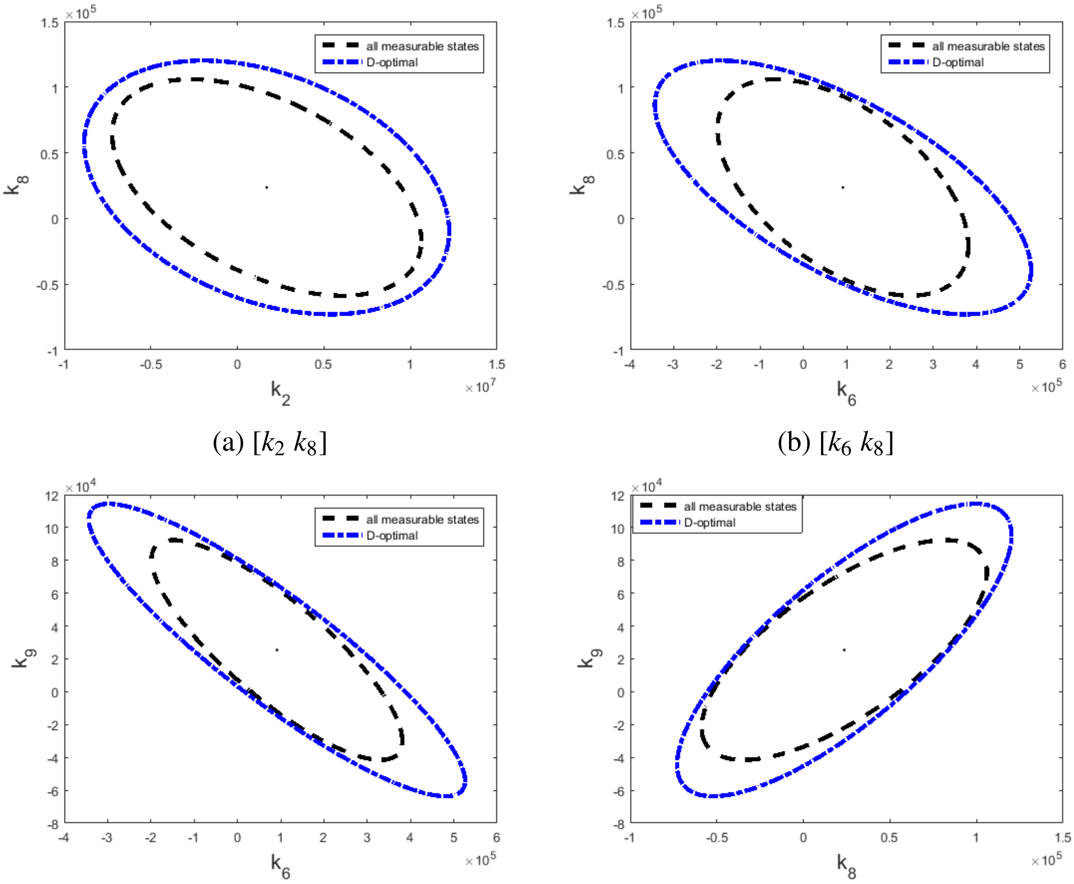
(d) $[k_8, k_9]$

Fig. 5.11 CIs under *D*-optimal and emprical sampling strategies. The point in the middle of the figure indicates the nominal parameter values.

selected for the measurement. However, the D-optimal design result reveals that only three state variables D, BD and FFA are important measurement targets while the other two state variables T and M are not important at all. To further examine these two design results, parameter CIs are compared in Figure 5.12. The blue ellipse corresponds to the situation when only three state variables D, BD and FFA are used as the measurement set. The black ellipse corresponds to the results by using all the five state variables as the measurement set. It can be observed that the D-optimal design are not very close to that with all the five measurable state variables selected, which indicates that state variables T and M are also important. The measurement of these two states can provide useful data for parameter estimation. Therefore, in this case the E-optimal design can generate more reliable results than the D-optimal design for the measurement set selection.

5.4 Summary

In this chapter, the model-based OED was investigated through two case studies of biochemical systems with individual experimental factor considered. Three key experimental factors are designed, i.e. the input intensity, the sampling profile and the measurement set selection. For each of the design, the OED formulations and numerical optimisation was discussed and their performance has been examined by the two exemplar studies. For the benchmark enzyme reaction system, the initial input conditions have been optimised so that the intended experiment can generate maximum data information. Then the sampling time design is applied. By using the Powell's method, the optimal sampling points have been determined. The importance of our sampling time design method is that it can decide the informative time regions within which all the data are relatively more informative than that from other time points. Therefore, it is useful in real industry as it allows an effective sampling strategy during experimentation. In the OED for the measurement set selection, the substrate S and



(a) $[k_2\ k_8]$

(b) $[k_6\ k_8]$

(c) $[k_6\ k_9]$

(d) $[k_8\ k_9]$

Fig. 5.12 Comparison of CIs for k_2, k_6, k_8, and k_9 under different strategies of measurement sets

desired product Q are suggested to be the two most valuable state variables which should be selected to do the measurement. In the experimental design for the enzymatic biodiesel production system, two typical sampling regions have been determined which can lead to smaller parameter estimation errors than the non-designed condition. By using the E-optimal design for the measurement set selection, all five measurable states are identified to be important, the measurement of which can provide useful data.

Chapter 6

Comprehensive Experimental Design - A Two-layer Iterative Strategy

In Chapter 5, we investigated the OED, particularly, on single factors which include input design, sampling selection and measurement set selection. The OED procedure is described through the case studies, from which it can be seen that the bound of parameter uncertainties can be decreased. With OED on individual factors, in each design, other experimental factors are fixed. The dependence between factors will affect the design results. Therefore, it is necessary to develop an efficient method which can optimise all design factors in an integrated framework and this will lead to more accurate parameter estimation than single factor design. Finding experimental designs which are optimal with respect to multiple factors can be accomplished by performing an integrated optimisation of the experimental degrees of freedom. In this chapter, the comprehensive experimental design will be investigated and applied to the two biochemical systems.

6.1 Integrated observation design of measurement set and sampling profile

The observation design mainly concerns the problem of measurement strategy during experimentation, which includes but is not limited to sampling time design and measurement set selection. The main purpose of integrating different measurement related factors into one observation design problem is to develop an observation design problem that is not only numerically simple to solve but also effective for data generation for parameter estimation. For observation design under given input condition, the dynamic response of the system is fixed, which means numerically the ODEs of the dynamic model only need to be solved once during the optimisation process. It is therefore possible to include different observation designs into a single OED problem.

The design of sampling time profile and the design on measurement set selection are usually handled separately, as described in Chapter 5. In each design, it is assumed that all the other experimental factors are specified. This single factor design may fail to give a satisfactory result to guide measurement data collection since the experimental factors in observation are related to each other in terms of providing information content. A more effective OED should put the multiple observation factors together into one integrated design. One option to consider both sampling time profile design and measurement set selection is to go through an iterative procedure to design the two experimental factors, in each iteration only one factor is optimised based on the predefined settings of the other one, and repeats until both factors are properly designed. This iterative procedure is not computationally efficient, also the dependent effects of the two measurement factors are still handled separately during the design.

When considering measurement set selection and sampling time design together, there are wider options available for both sets of design factors within the design domain. As can be seen in Section 5.1, the sampling time design problem can be represented in a similar form to the measurement set selection design formulation. When the choice of sampling points for one state variable can be independent from the time scheduling of another state variable, the measurement set selection and the sampling time design problem can be combined together, and that leads to the integrated observation design, the formulation of which can be represented as:

$$\begin{aligned} \boldsymbol{\zeta} &= \left\{ \begin{array}{cccc} t_1 & t_2 & \cdots & t_{N\times n} \\ \omega_1 & \omega_2 & \cdots & \omega_{N\times n} \end{array} \right\} \\ \boldsymbol{\zeta}^* &= \underset{\boldsymbol{\omega}\in\boldsymbol{\Omega}}{\arg\min}\, \psi\left(\left(\sum_{i=1}^{N\times n} \frac{1}{\sigma_i^2}\omega_i \mathbf{S}(t_i)^T \mathbf{S}(t_i)\right)^{-1}\right) \\ & s.t. \quad \omega_i \in \{0,1\}, \mathbf{1}^T\boldsymbol{\omega} = N_{sp} \end{aligned} \tag{6.1}$$

where the number of the integrated weighting factors is extended to $N \times n$. Each weighting factor stands for the importance of one measurable state variable at a particular time point. Equation (6.1) is an integer programming problem which can be solved by exhaustive search with a relatively small number of available time points. For OED problems that contain a large number of measurable state variables and sampling time points, problem (6.1) can be further relaxed to an approximate continuous optimisation problem which is given as follows:

$$
\begin{aligned}
\boldsymbol{\zeta} &= \left\{ \begin{array}{cccc} t_1 & t_2 & \cdots & t_{N\times n} \\ \omega_1 & \omega_2 & \cdots & \omega_{N\times n} \end{array} \right\} \\
\boldsymbol{\zeta}^* &= \arg\min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}} \psi\left(\left(\sum_{i=1}^{N\times n} \frac{1}{\sigma_i^2}\omega_i \mathbf{S}(t_i)^T \mathbf{S}(t_i)\right)^{-1}\right) \\
& s.t. \quad \sum_{i=1}^{N\times n} \omega_i = 1, \omega_i \geq 0
\end{aligned}
\tag{6.2}
$$

where the weighting term ω_i is relaxed to a continuous variable taking values between [0, 1]. In this way, the optimal solution provides a lower bound for the original integer optimisation problem. At each time point the FIM is a positive definite matrix. Therefore, the continuous optimisation problem can be converted into a convex optimisation problem by employing proper scalar design criteria. For instance, taking the D-optimal design criterion, the observation design problem can be easily transformed into a convex optimisation problem that can be solved by local optimisation methods such as the Powell's conjugate direction method or the interior-point method. When the A-optimal or E-optimal design criterion is applied, Problem (6.2) can be transferred into an equivalent semi-definite programming (SDP) problem.

Taking the E-optimal design criterion, the observation design formulation is written as follows:

$$
\begin{aligned}
& min \ -t \\
& s.t. \quad \sum_{i=1}^{n\times N} \frac{1}{\sigma_i^2}\omega_i \mathbf{S}_i^T \mathbf{S}_i \succ t\mathbf{I} \\
& \omega_i \succ 0, \forall i; \ \mathbf{1}^T \boldsymbol{\omega} = 1
\end{aligned}
\tag{6.3}
$$

The optimisation problem in (6.3) can be conveniently solved by available computational tools such as the '*SeDuMi*' software (Sturm, 1999). When a gradient-based optimisation method is used to solve this problem, the derivative of the objective function over the weights is much easier to calculate than the direct derivative over time and over the state variables.

With the proposed OED problem in 6.2, the sampling profile and the measurement set can be determined simultaneously.

In this work, we will compare the observation design approach with two sequential design methods, and the empirical sampling without optimisation design. The three experimental design strategies are listed in Table 6.1, in which Strategy 1 is a sequential method starting from measurement set design followed by sampling profile design; Strategy 2 is also sequential method but starting from sampling profile design followed by measurement set selection; Strategy 3 is the proposed approach that designs the sampling points and measurement set simultaneously. With the two sequential methods, the sampling time points are the same to those selected state variables. For the proposed Strategy 3, the sampling points for each selected state can be different.

Table 6.1 Three observation design strategies

OED methods	Design procedures
Strategy 1	Sequential: measurement set $\longrightarrow$ sampling time profile
Strategy 2	Sequential: sampling time profile $\longrightarrow$ measurement set
Strategy 3	Simultaneous: measurement set & sampling profile

6.2 Iterative double layer design of observation and input

In a systematic experimental design, those major experimental settings such as input perturbations and measurement strategy should be considered in an integrated design framework. This integrated optimisation problem can be handled through a process where the input design and the observation design are solved sequentially and iteratively until the satisfactory result is obtained. The input design problem can be formulated as a complex non-convex optimisation problem as discussed in Section 5.1, while the observation design problem is usually treated as a convex optimisation problem as described in Section 5.1 and Section 6.1. As such, there is no simple solution for this multi-factor optimisation problem.

In this work, we propose an iterative double-layer procedure, as illustrated in Fig.6.1, to design the experimental factors for both the input and the observation. The design of inputs is arranged in the upper layer, and the observation design is put in the lower layer. Due to the non-convex nature of the input design problem, a modern heuristic method, particle swarm algorithm (PSA) (Kennedy, 2011), is chosen to obtain the optimal solution. This method is a population-based optimisation algorithm which can solve a variety of hard problems with fast convergent rate. With this algorithm, only a few parameters need to be tuned and no derivative calculations are required, making the algorithm attractive from the implementation point of view. The basic PSA method is based on a population of s particles that represents solutions of the optimisation problem. Each particle is associated with a position x and a velocity v, which denote its position and movement through the search space. The position and velocity of a particle can be dynamically adjusted through an iterative process according to the objective function values at particle positions. At the generation k, the new position x_i^{k+1} of the i-th particle is computed by adding to the old position x_i^k a velocity vector v_i^{k+1}:

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{6.4}$$

The velocity vector of the i-th particle is updated by

$$v_i^{k+1} = \omega \cdot v_i^k + \alpha_1 \cdot r_1 \cdot \left(pbest_i^k - x_i^k\right) + \alpha_2 \cdot r_2 \cdot \left(gbest^k - x_i^k\right) \tag{6.5}$$

where ω, α_1 and α_2 are the inertia parameter, the cognition parameter and the social parameter, respectively. r_1 and r_2 are numbers randomly chosen in the range of 0 to 1. $pbest_i^k$ is the best position of the i-th particle up to the k-th generation, and $gbest^k$ is the best position of the k-th generation among all particles, which can be determined by:

$$gbest^k = \underset{z \in x_1^k, x_2^k, \cdots, x_s^k}{\arg\min} f(z) \tag{6.6}$$

where $f(\cdot)$ is the objective function. The detailed PSA implementation for the input design is described as follows.

Algorithm 6.1

1. Choose a population size s and the iteration number n_{tol}. Initialise the swarm positions x_1^0, x_2^0, $\cdots$, x_0^s and their velocities v_1^0, v_2^0, $\cdots$, v_0^s.
2. Let $pbest_i^0 = x_i^0, i = 1, 2, \cdot, s$, and determine $gbest^0$ using equation (6.6), let $k = 0$.
3. Set $gbest^{k+1} = gbest^k$. For every particle i, do:
   - Check the constraint of x_i^k, make sure each particle is within the bound.
   - If $f(x_i^k) \leq f(pbest_i^k)$, then update the best position of the i-th particle, $pbest_i^{k+1} = x_i^k$; if $f(pbest_i^{k+1}) \leq f(gbest^{k+1})$, then update the best position at current generation, $gbest^{k+1} = pbest_i^{k+1}$; otherwise, set $pbest_i^{k+1} = pbest_i^k$.
4. Compute x_i^{k+1} and v_i^{k+1} for each particle using equations (6.4) and (6.5).
5. Stop when $k = n_{tol}$. Otherwise, increment k by one and go to step 3.

In this double-layer structure, the inputs are firstly determined by a running pre-defined number of iterations of the PSA method, based on which the combined observation design problem is solved at the lower layer through the Powell's conjugate direction method (Fletcher and Powell, 1963). The designed observation strategy is then used in the next iteration for input design until the optimal solution for both the input and the observation is obtained.

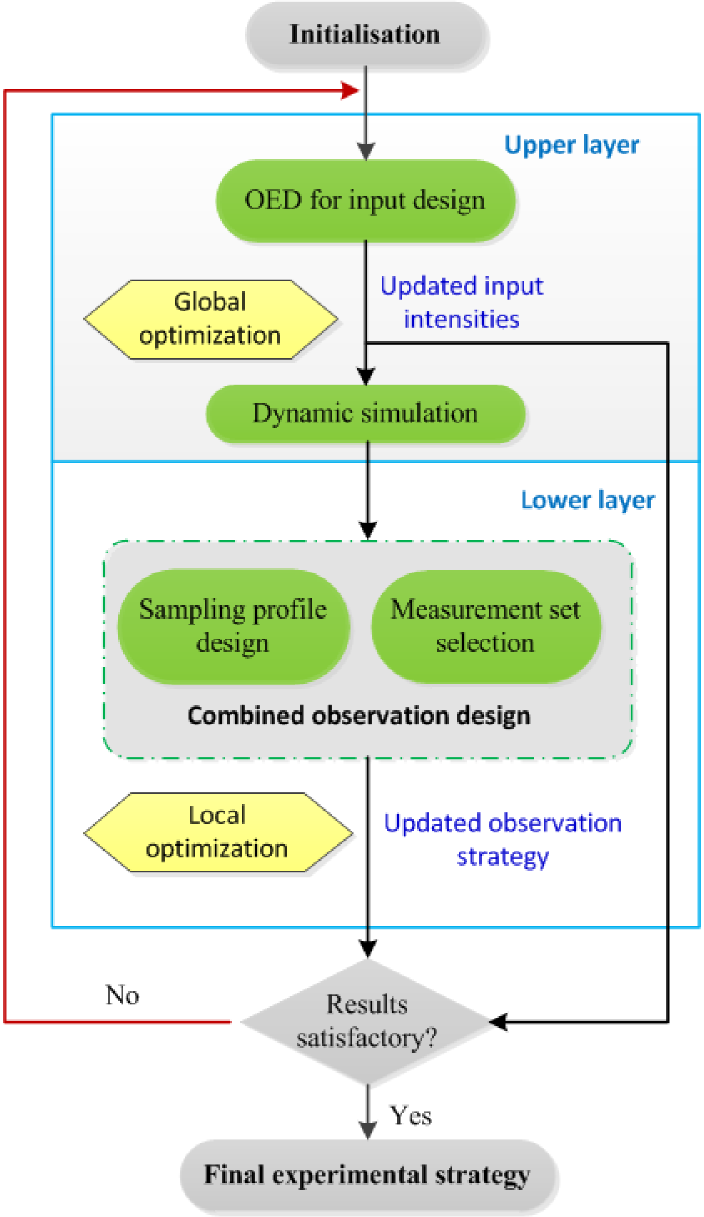


Fig. 6.1 Iterative double-layer design for both input and observation factors

At the lower layer the convex optimisation problem of observation design under the given input conditions can be solved. At the upper-layer, employing stochastic searching largely increases the chance of finding a global solution. This is a clear advantage over the traditional local numerical algorithms which most likely only lead to local optimum. For a complex design problem including both input design and observation design, it is also computationally more efficient to put the observation design at the lower layer since this is a convex optimisation problem that is relatively easy to solve. The procedure of the iterative double-layer optimisation to solve the integrated OED problem is given as follows.

Algorithm 6.2

1. Initialize the objective function $g(x,y)$, where x and y denote the input variables and observation strategy, respectively. Set the stopping tolerance $\delta_{tol} \geq 0$.

2. Let the iteration number $l = 0$, calculate y^0_{best} based on x_{nom} (x_{nom} is a vector of the nominal values of the input variables) using Powell's method. Then determine x^0_{best} for $g(x, y^0_{best})$ using Algorithm 6.2.

3. For generation l, determine x^l_{best} for the objective function $g(x, y^{l-1}_{best})$ in the upper layer using PSA method described in Algorithm 6.2, then calculate y^l_{best} for $g(x^l_{best}, y)$ in the lower layer using Powell's method.

4. If $\left|g^{l+1} - g^l\right| \leq \delta_{tol}$, then stop the optimisation process. Otherwise, increment l by one and go back to step 3.

Using this iterative double-layer numerical strategy, the input design and the observation design problems can be integrated into one optimisation framework. Compared with the sequential design process where each OED problem is solved only once, the proposed method enables the update of the input variables and the observation strategies during each iteration of the optimisation process. In this way, the design order of input and observation factors does not need to be considered.

6.3 Case study 1: Comprehensive design for enzyme reaction system

6.3.1 Integrated observation design

For the enzyme reaction system, we first conduct the OED on combined observations. The estimation parameters are still k_2, k_{-3} and k_5W. The three experimental design strategies shown in Table 6.1 are applied to this model by using the D-optimal design criterion. The objective in observation design is to determine important state variables that potentially have significant contributions to parameter estimation, and to obtain 100 sampling points for these important state variables which will lead to the most informative data for parameter estimation. The input conditions and parameter values are given in Table 4.2 and Table 4.3. In addition, the Gaussian white noise with 10% relative error and 0.0001 absolute error are added to the simulation data. The simulation results with the three OED strategies and the non-designed equally spaced sampling are shown in Table 6.2.

We start from observation design by taking the nominal parameter values and the initial conditions in Table 4.3 and Table 4.2. The design objectives are: (i) to select measurement state variables; and (ii) to locate 100 sampling time points for the selected state variables, which will lead to the most informative data set for parameter estimation. In the simulation, zero-mean Gaussian white noises are assumed for measurement errors with 0.1 relative and 0:001 absolute contributions. Three different design strategies, as shown in Table 6.1, are compared during the simulation. A sequential design procedure is taken as Strategy 1 and Strategy 2, where the former starts with the measurement set selection followed by the sampling profile design, and the latter starts from the sampling profile design followed by the measurement set selection. In the proposed integrated design, named as Strategy 3, the design of the two tasks are combined into one single optimisation problem and the solutions for both can be obtained simultaneously. The D-optimal design criterion is employed in all the three OED methods.

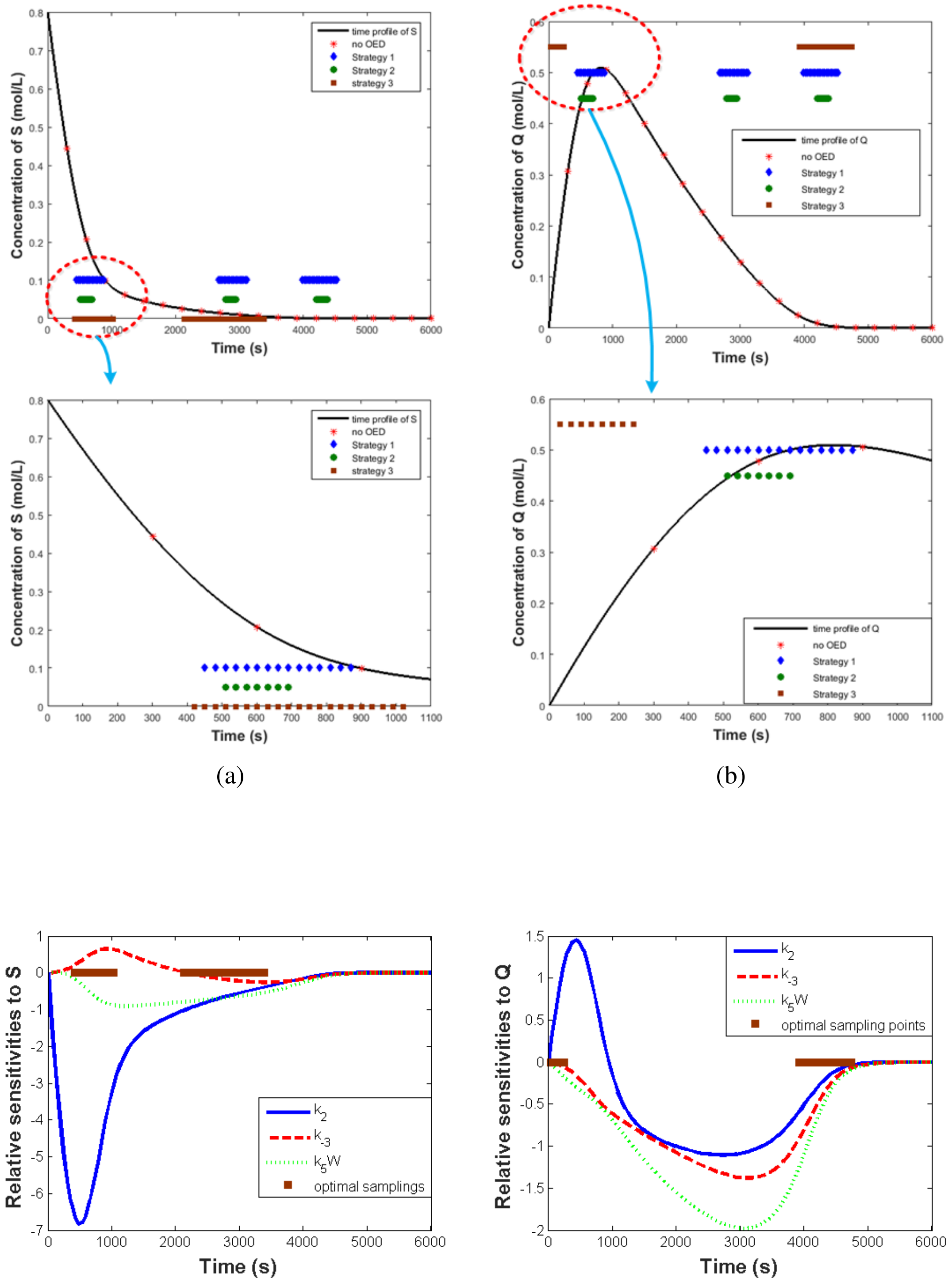
The design results of the three observation strategies and also the default setting without any OED are listed in Table 6.2. When no OED is employed, all the 5 measurable states are taken into account, and the same equally-spaced sampling rule is applied to all the 5 states, i.e., 20 sampling points for each state with sampling period of 300 seconds. For OED with Strategy 1, the two variables, S and Q, are firstly selected to form the measurement set, then the sampling profile design is performed to S, Q, which gives 3 sampling regions. In Strategy 2, the sampling design is made first to all the 5 states and 3 sampling regions are found. Then using the designed sampling profile, the measurement set is selected which in fact includes two states, S and Q. Instead of taking these two variables, all five measurable variables are included for the measurement set otherwise the total number of data will be reduced. With the proposed Strategy 3, the total number of 100 sampling points are 'allocated' to S and Q after the optimal design. It can be seen that both Strategy 1 and Strategy 3 select S and Q as the most important measurement variables although the sampling profiles are different.

Table 6.2 Observation design results for enzyme reaction system

Methods	Selected measurement states	Sampling time points (second)	D-value
Non-designed condition	S, P, N, Q, R	[300:300:6000]	4.03e-6
Strategy 1	S, Q	[450:30:870], [2670:30:3120], [3390:30:4530]	1.83e-5
Strategy 2	S, Q	[510:30:690], [2790:30:2940], [4200:30:4380]	1.83e-5
Strategy 3	S	[420:30:1020], [2130:30:3390]	7.12e-4
	Q	[30:30:240], [3930:30:4740]	

For Strategy 1, Strategy 2 and the no-OED cases, all (or selected) variables have the same sampling profile. Only with Strategy 3, the sampling profiles for each selected variable can be different. Taking S and Q as the selected state variables, the sampling points distribution from different experimental strategies are shown in Figure 6.2. With the sequential design of Strategy 1 and Strategy 2, three sampling regions are recommended at different reaction stages, mostly corresponding to where the variables or local sensitivities have large changes. The sampling regions of Strategy 2 are narrower compared to Strategy 1. This is because there are five measurement variables in Strategy 2 and only two variables in Strategy 1 at the design of sampling profile. Using the proposed Strategy 3, two sampling regions are found for S and two for Q, respectively, covering a wider range of the reaction process. Within each sampling region, consecutive measurement points are recommended by the design result which indicates that measurement within those selected sampling regions can potentially provide informative data.

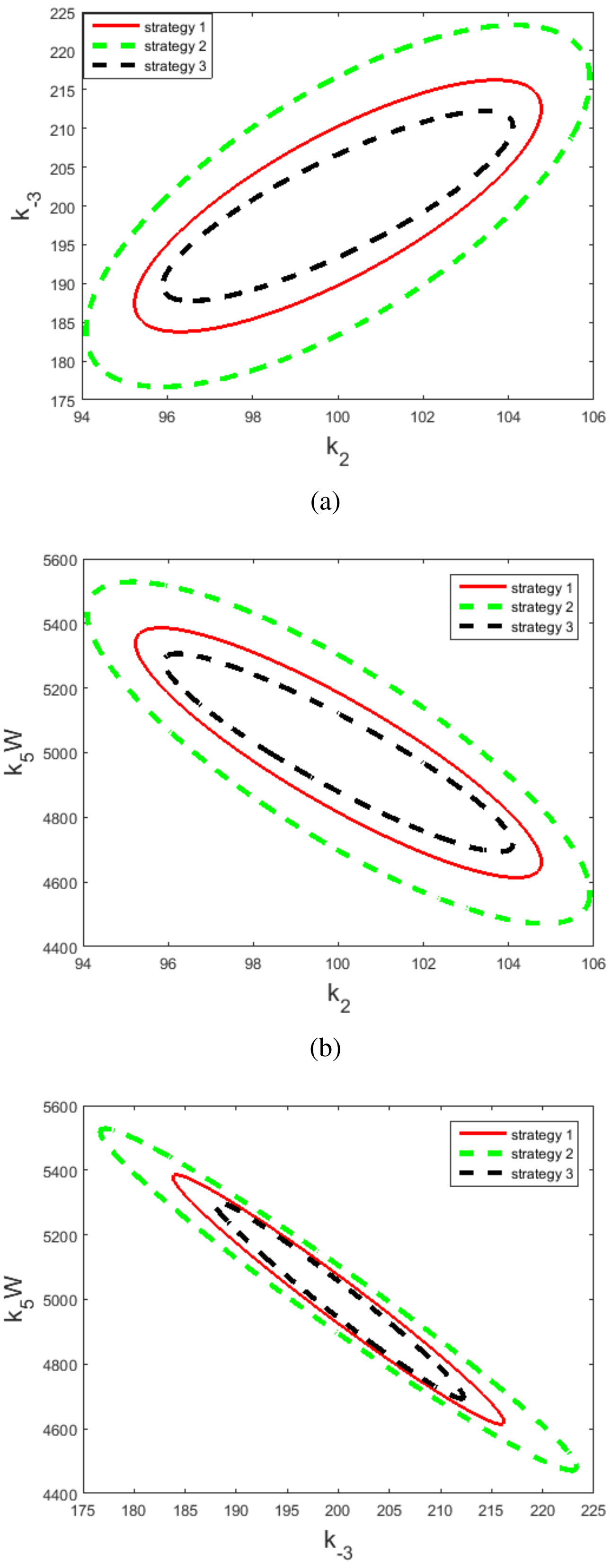
The CIs of the three key parameters in pairs are compared in Figure 6.3. According to the Cramer-Rao inequality, a smaller CI region corresponds to smaller lower bounds for parameter estimation errors, therefore a better estimation quality can possibly be obtained. In this simulation, the design Strategy 1 shows smaller CIs than Strategy 2, which suggests that in the sequential design, the measurement set should be selected prior to the sampling time design. The proposed integrative design, Strategy 3, achieves the smallest CIs among the three methods due to the fact that the two observation factors are simultaneously determined during the OED. The OED of observation provides a useful insight that the sampling points should be taken at several critical regions that correspond to large parameter sensitivities or large change rates in key variables, not necessarily equally spaced as in a traditional way.



(a) (b)

(c) (d)

Fig. 6.2 Sampling profiles of S and Q with different strategies



(a)

(b)

(c)

Fig. 6.3 Comparison of CIs for kinetic parameters under different observation design strategies

6.3.2 Iterative two-layer design of observation and input

In this section, input and observation variables are designed together through the proposed iterative double-layer OED strategy as shown in Figure 6.1. The results are compared to another iterative OED, in which the observations are designed sequentially using Strategy 1, as discussed in Section 6.3.1. In both cases, the iteration number for the whole optimisation is set to be 100, and the typical run time of each complete design is around 1.5 hour on a personal computer with i5-2400 CPU and 4 GB memory. The designed results are shown in Table 3. By considering both the input and observation factors, S and Q are selected for the measurement set with Strategy 1 used at the lower layer, which is consistent with the observation design results in Section 6.3.1. The state of N is also found important for measurement set when Strategy 3 is used in the observation design. The CIs of the selected key parameter pairs are shown in Figure 6.4. Again, it can be seen that using similar computational time, the results from the proposed OED method provides smaller parameter estimation uncertainty bound (Figure 6.4b), reduced correlation between parameter pairs (Figure 6.4a and 6.4c), compared with the method following sequential observation design.

Table 6.3 Input and observation experimental design results with different numerical strategies

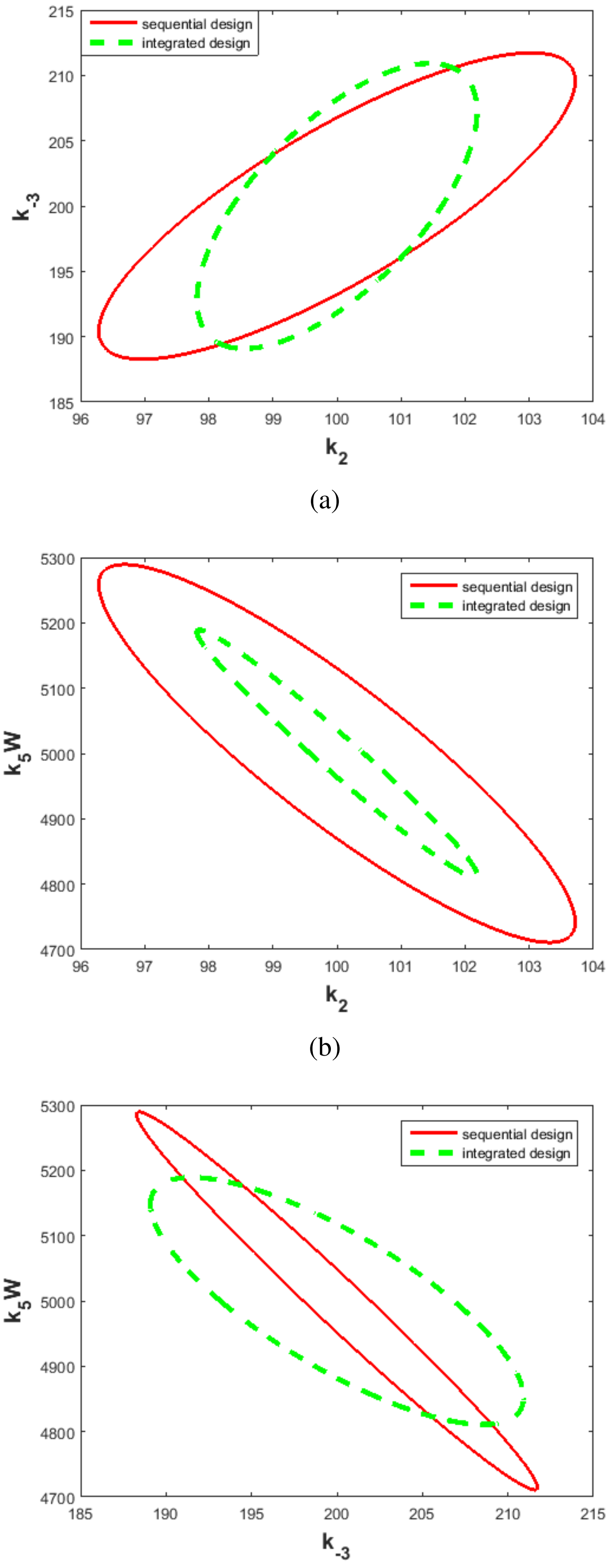
Two-layer OED	$[S_0, E_0, N_0]$ $(mol \cdot L^{-1})$	Measurement set	Sampling points (second)	D-values
Sequential design	S_0: 0.74 E_0: 1.52e-5 N_0: 1	S Q	[420:30:1020], [2130:30:3390] [30:30:240], [3930:30:4740]	0.0019
Iterative two-layer design	S_0: 1 E_0: 5.64e-6 N_0: 0.13	S Q N	[4590:30:5760] [5280:30:6000] [390:30:1410]	0.0027

One should note that using the proposed observation design or the iterative double-layer design strategy, several critical regions are suggested for sampling rather than the equally spaced sampling over the whole process time range. The latter has been widely used in chemical engineering. Taking uniform sampling at the very early stage of modelling and design would be useful, where model information is limited and parameter values contain large uncertainties. At a later stage when more modelling knowledge is available, the regional sampling strategy by OED will provide more useful data information for parameter estimation.

6.4 Case study 2: Integrated observation design for enzymatic biodiesel production system

As discussed in Chapter 5.3, the input signal for this biodiesel production system has been predefined following the production requirement. Only the observation will be designed through OED. Taking the four most important parameters, k_2 k_6 , k_8 and k_9, into the parameter estimation scheme, OED has been applied to determine the best observation strategy which includes the most valuable measurement state variables and the best sampling time points for the measurement states. Considering the reality of experimentation, the minimal sampling time interval between two sampling points is set to be 5 minutes. In non-designed settings, 28 empirical sampling points were selected for all five measurable state variables which are T, D, M, BD and FFA. In order to compare the OED design results to real experimentation work, the number of sampling points in this simulation is chosen to be 140 (28×5) in total. Three different experimental strategies in Table 6.1 are tested, the results of which are shown in Table 6.4 and Figure 6.5.

It can be seen that Strategy 1 and Strategy 2 make exactly the same result on the sampling time profile and measurement set. in both designs all five measurable state variables are selected for measurement. By using the integrated observation design Strategy 3, only four

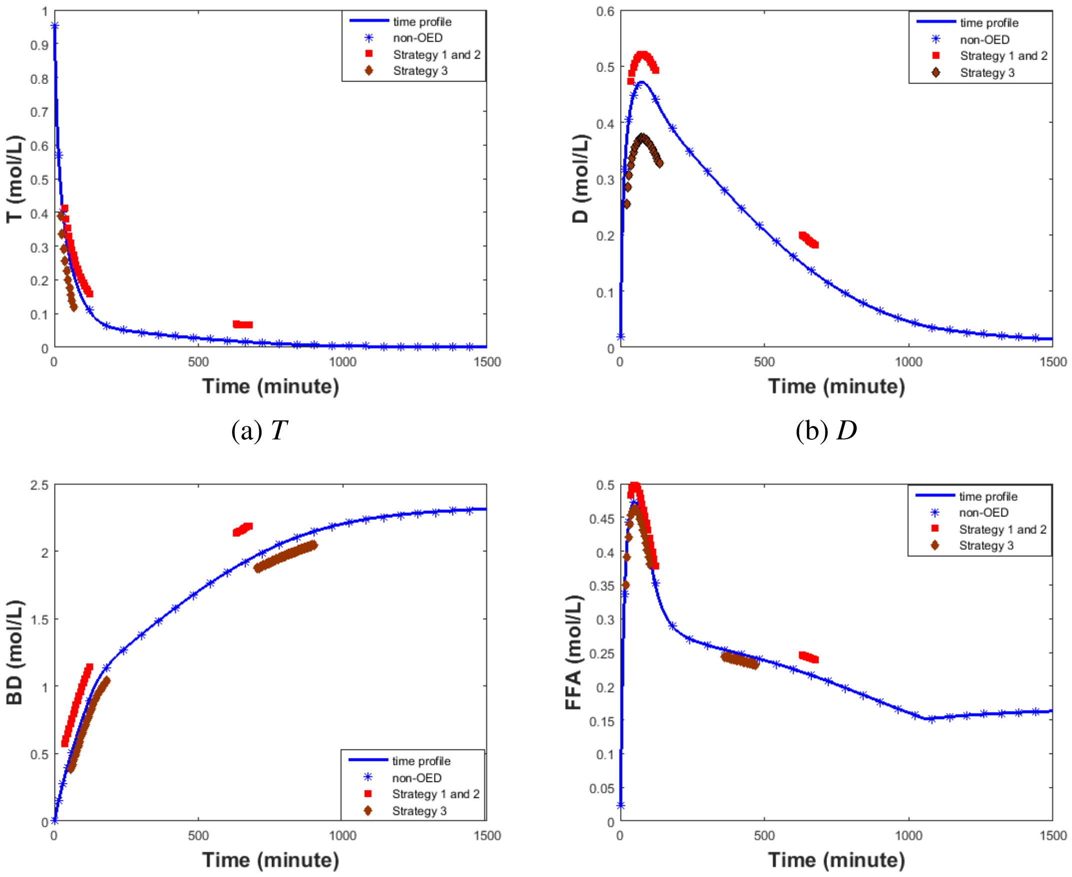


(a)

(b)

(c)

Fig. 6.4 Comparison of parameter uncertainties with different numerical strategies



(a) *T*

(b) *D*

(c) *BD*

(d) *FFA*

Fig. 6.5 Sampling time points on the selected state variables

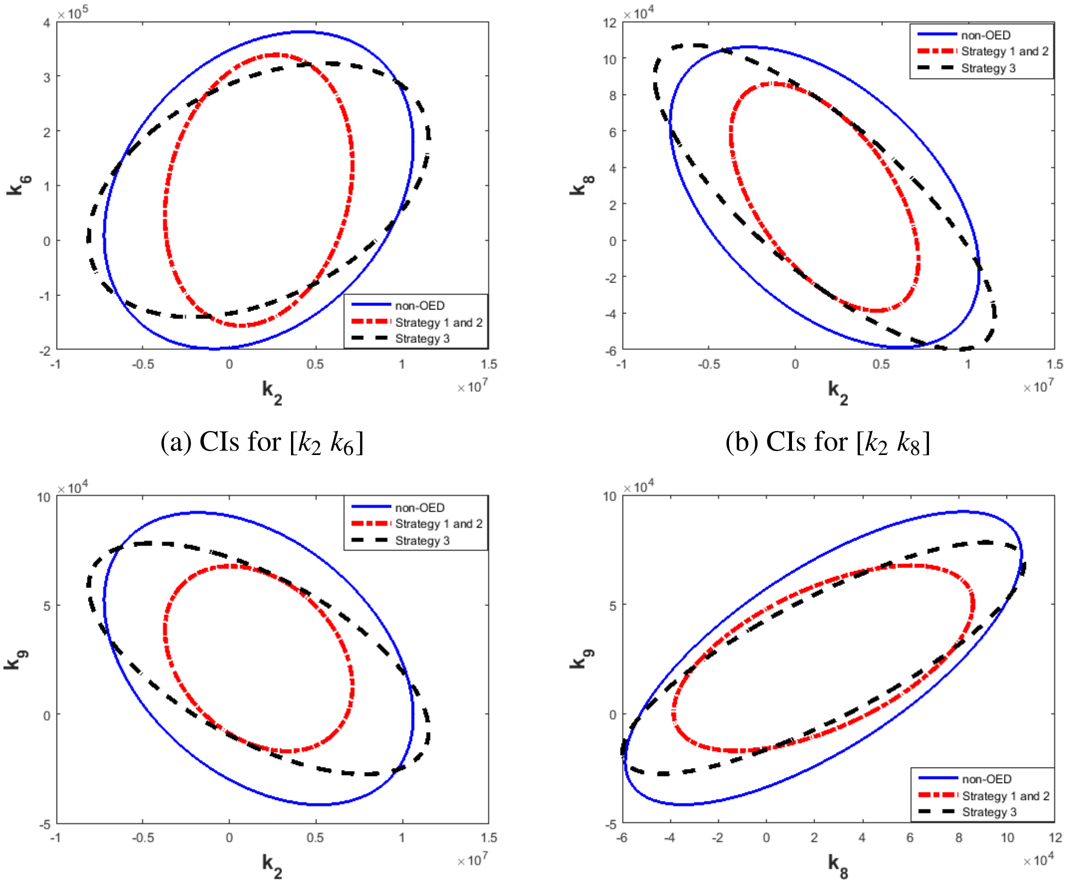
Table 6.4 Design results of different OED strategies

	Measurement state variables	Sampling profile (unit: minute)	D-value
no-OED	T, D, M, BD, FFA	[15:15:60], [120:60:1440]	5.39e-41
Strategy 1	T, D, M, BD, FFA	[35:5:120], [629:5:674]	3.96e-40
Strategy 2	T, D, M, BD, FFA	[35:5:120], [629:5:674]	3.96e-40
Strategy 3	T	[21::5:66]	1.57e-39
	D	[20:5:135]	
	BD	[57:5:182], [704:5:899]	
	FFA	[17:5:102], [361:5:466]	

state variables need to be selected, which are T, D, BD and FFA. The state variable M can be ignored during the measurement.

In terms of the sampling time profile, it can be found that (see Table 6.4 and Figure 6.5) two sampling time regions are suggested for the measurement from the sequential design (Strategy 1 and 2). One is at the early stage of the reaction (the first two hours) and the other is in the middle of the reaction (between 600 to 700 minutes). The sampling time from Strategy 3 is a bit different from the sequential design. From Figure 6.5a and 6.5b, only one sampling time region is required for state variables T and D. The samples are taken nearly in the first hour for T and in the first 2 hours for D. This is reasonable because in the kinetic reaction T is taken reaction prior to D. For state variables BD and FFA, two sampling regions are suggested which are at the early stage and middle stage of the reaction. Strategy 3 suggests to take more samples for BD and FFA than in Strategy 1 and 2. Also, the largest number of samples for BD among the four selected states is suggested which indicates that BD is the most important variable for the measurement.

The observation design results are further assessed by comparing the CIs of key parameter pairs in 6.6. It can be seen that CIs of all OEDs are smaller than the scenario without OED, and the proposed Strategy 3 achieves the smallest CIs among all OEDs.



(a) CIs for [k_2 k_6]

(b) CIs for [k_2 k_8]

(c) CIs for [k_2 k_9]

(d) CIs for [k_8 k_9]

Fig. 6.6 Comparison of CIs for different strategies

6.5 Summary

This chapter illustrated the optimal experimental design with consideration of several different objectives simultaneously. The sampling time design and measurement set design are combined together into one observation design problem. Through simple transformation the observation design can be reduced to an easily solved convex optimisation problem. Then we proposed a double-layer optimisation algorithm to solve the integrated experimental design which combines input factor design and observation design. The input design which is a non-convex optimisation problem is solved by using a PSA algorithm in the upper layer and the convex observation design problem is solved by using Powell's method in the lower layer. This numerical strategy has been proved to be superior to the sequential design procedure through its application to the integrated design of the case study biochemical systems. The method we proposed is general and can be easily extended to the application of the OED for other biochemical systems.

Chapter 7

Conclusions and Future Perspectives

This chapter provides summary and concluding remarks for the work presented in this thesis and suggestions for potential future work.

7.1 Conclusions

This thesis was concerned with model-based OED for chemical engineering systems with the aim of developing advanced experimental design approaches to reduce parameter estimation uncertainties. The main objective was to propose efficient design methods and numerical strategies that could deal with OED problems for general dynamic models. The thesis includes three main parts. The first part is a comprehensive study of existing OED techniques and related analysis tools presented in Chapter 2 and Chapter 3. The challenges of data-based modelling was discussed and the usefulness of OED for effective modelling was highlighted. Limitations and problems of existing OED methods were illustrated and some possible research aspects were pointed out. In the second part, two biochemical systems, the enzyme reaction system and the enzymatic biodiesel production system, were described, and preliminary analysis for these two models was conducted to facilitate further OED development. The enzyme reaction system was proposed which can represent a typical class of enzyme kinetically controlled synthesis process. The investigation of this type of system is of particular importance because the proposed OED methods for this system can be easily extended to other systems with similar reaction schemes. The enzymatic biodiesel production model represents a lab-scale real transesterification process. The investigation of this model in the OED context is important as it allows the performance analysis of the proposed OED methods on real problem. The third part is the OED development and application to these two biochemical systems presented from Chapter 5-6. General formulations for OED problems by considering different design factors were proposed in Chapter 5. Also, the effectiveness of the developed methods were demonstrated through their applications to those two biochemical systems. Finally, a novel numerical strategy for the OED problem with integrated design factors was created in Chapter 6. The observation design and its combination with input design have been investigated. This chapter summarises the main conclusions of this thesis and suggestions for future work.

Performing OED for complex models containing a considerable number of parameters requires an a priori choice to be made about the number of parameters that will be taken into account for the design. Finding optimal experiments that would result in data of such high quality that all parameters can be estimated is very difficult, because some parameters might not be identifiable. The parameter sensitivity and model identifiability have been analysed in Chapter 4. The generating series approach, combined with identifiability tableau, has been applied to check whether the unknown model parameters are structurally identifiable. Then the practical identifiability analysis has been conducted, in which the PCA based method, the collinearity index method and the orthogonalised sensitivity analysis method have been compared in order to determine important parameters which are influential to model outputs. This preliminary analysis is essential to further OED development.

The general formulation and numerical optimisation strategies for model-based experimental design by considering different design factors have been developed in Chapter 5. OED for input design has been formulated as a non-convex optimisation problem. Various numerical algorithms are compared through the case study of an enzyme reaction system. From the simulation result, it is found that differential evolutionary algorithm can lead to better results, from which the generated data are more informative for parameter estimation. The sampling time design and measurement set design have been investigated in the biochemical modelling context. The sampling time selection is treated as a discrete optimisation problem based on the fact that parametric sensitivities at adjacent points are close to each other. The resulting discrete optimisation problem can then be relaxed into a continuous convex optimisation problem which can be solved efficiently.

An integrated observation design that considers both measurement set selection and sampling time scheduling has been proposed in Chapter 6. By approximating available sampling points in priori, the problem formulation for sampling time design can be expressed in a similar form as measurement set design. Therefore, these two design tasks can be combined together as a single optimisation problem, which is further relaxed to a convex optimisation problem that can be conveniently solved using local optimisation methods. Through the case studies based on an enzyme reaction model and a kinetic model developed for a lab-scale enzymatic biodiesel production process, the effectiveness of the integrated observation design over traditional sequential design strategy has been examined. In both case studies, the lower bounds for parameter estimation errors can be reduced through the observation design. Another advantage of this proposed method is that it can automatically choose the number and position of measurement points for each measurable state variable rather than measuring all state variables using the same sampling profile.

The OED by considering various experimental degrees of freedom simultaneously is not a trivial task. Very few works have been focused on the integrated design problem.

When using a classic non-linear optimisation algorithm there is a high probability that only a local minimum can be found. A new double-layer optimisation strategy was proposed in Chapter 6 to solve the integrated experimental design problem. It combines input design and observation design together. The input design which is a non-convex problem is solved in the upper layer by a modified PSA method. The observation design, which is relaxed as a convex problem, is solved in the lower layer by the interior-point method. The updated input values and observation strategies are transferred between these two layers which could lead to faster convergence. In this way the integrated OED optimisation problem can be solved faster in many situations since no repeated sensitivity analysis is needed to optimise the observation strategy. The proposed strategy has been applied to the OED of the enzyme reaction system. The simulation result clearly shows that the global method outperforms the local optimisation techniques, and the algorithm is able to locate the global optimum in a consistent way. The designed experiment can generate more informative data than by considering single design factor only.

7.2 Open challenges and future perspectives

Model-based OED is a research area of growing interest in recent years especially for modelling of complex biochemical and biological networks. In this thesis, new development in OED has been explored mainly for biochemical systems modelling. To expand the development and methods to wider application areas, there are still many open questions and problems to be addressed. Some future perspectives on OED are discussed in the following.

1. The parameter identifiability analysis methods are still limited and require further investigation. Current parameter selection techniques could only determine which parameters should be estimated based on the available experimental data. In other words, it can only determine the estimability of model parameters under fixed experi-

mental conditions. However, modellers may be interested in the model behaviour at operation points that are different from those where measurements are available. The OED can provide data with higher information content and can increase the probability that other parameters will become identifiable. Therefore, there is a need for the development of parameter identifiability techniques that combine with OED so as to update identifiable parameter subsets and make more precise model predictions under various experimental conditions.

2. An important further work is to develop novel RED approaches that consider various design factors simultaneously into one optimal optimisation framework for OED problems. Efficient methods are required that can achieve a good compromise between computational efficiency and accuracy in parameter uncertainty representation.

3. The use of multi-objective experimental design has great potential for designing realistic measurement campaigns. Incorporating more objectives like model prediction uncertainty, experimental cost and other more practice oriented objectives could also prove beneficial.

4. The development of efficient OED methods to deal with different kinds of models is another interesting research aspect, for instance, the OED for stochastic models. It is necessary to develop new methods to formulate the OED for uncertain models and propose efficient ways to measure data information. Similarly, the OED by considering non-Gaussian type noise is also a interesting research topic.

References

Alaña, J. E. and Theodoropoulos, C. (2012). Optimal spatial sampling scheme for parameter estimation of nonlinear distributed parameter systems. *Computers & Chemical Engineering*, 45:38–49.

Alberton, A. L., Schwaab, M., Lobão, M. W. N., and Pinto, J. C. (2011). Experimental design for the joint model discrimination and precise parameter estimation through information measures. *Chemical Engineering Science*, 66(9):1940–1952.

Alberton, K. P., Alberton, A. L., Di Maggio, J. A., Díaz, M. S., and Secchi, A. R. (2013). Accelerating the parameters identifiability procedure: set by set selection. *Computers & Chemical Engineering*, 55:181–197.

Anderson-Cook, C. M. (2007). *Functional Approach to Optimal Experimental Design*. Taylor & Francis.

Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A., and Blom, J. G. (2009). Systems biology: parameter estimation for biochemical models. *The FEBS Journal*, 276(4):886–902.

Asprey, S. and Macchietto, S. (2000). Statistical tools for optimal dynamic model building. *Computers & Chemical Engineering*, 24(2-7):1261–1267.

Asyali, M. H. (2010). Design of optimal sampling times for pharmacokinetic trials via spline approximation. *Turkish Journal of Electrical Engineering & Computer Sciences*, 18(6):1019–1030.

Ataíde, F. and Hitzmann, B. (2009). When is optimal experimental design advantageous for the analysis of Michaelis–Menten kinetics? *Chemometrics and Intelligent Laboratory Systems*, 99(1):9–18.

Atherton, R. W., Schainker, R. B., and Ducot, E. R. (1975). On the statistical sensitivity analysis of models for chemical kinetics. *AIChE Journal*, 21(3):441–448.

Atkinson, A. (1996). The usefulness of optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):59–76.

Atkinson, A. C. (2008). Dt-optimum designs for model discrimination and parameter estimation. *Journal of Statistical planning and Inference*, 138(1):56–64.

Atkinson, A. C. and Fedorov, V. V. (1975a). The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70.

Atkinson, A. C. and Fedorov, V. V. (1975b). Optimal design: Experiments for discriminating between several models. *Biometrika*, 62(2):289–303.

Audoly, S., Bellu, G., D'Angio, L., Saccomani, M. P., and Cobelli, C. (2001). Global identifiability of nonlinear models of biological systems. *IEEE Transactions on Biomedical Engineering*, 48(1):55–65.

Balsa-Canto, E., Alonso, A. A., and Banga, J. R. (1998). Dynamic optimization of bioprocesses: deterministic and stochastic strategies. *ACoFop IV (Automatic Control of Food and Biological Processes), Göteborg, Sweden*, pages 21–23.

Balsa-Canto, E., Alonso, A. A., and Banga, J. R. (2008). Computational procedures for optimal experimental design in biological systems. *IET Systems Biology*, 2(4):163–172.

Balsa-Canto, E., Alonso, A. A., and Banga, J. R. (2010). An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Systems Biology*, 4(1):11.

Balsa-Canto, E., Rodriguez-Fernandez, M., and Banga, J. R. (2007). Optimal design of dynamic experiments for improved estimation of kinetic parameters of thermal degradation. *Journal of Food Engineering*, 82(2):178–188.

Baltes, M., Schneider, R., Sturm, C., and Reuss, M. (1994). Optimal experimental design for parameter estimation in unstructured growth models. *Biotechnology Progress*, 10(5):480–488.

Banga, J. R., Balsa-Canto, E., Moles, C. G., and Alonso, A. A. (2005). Dynamic optimization of bioprocesses: Efficient and robust numerical strategies. *Journal of Biotechnology*, 117(4):407–419.

Banga, J. R., Versyck, K. J., and Van Impe, J. F. (2002). Computation of optimal identification experiments for nonlinear dynamic process models: A stochastic global optimization approach. *Industrial & Engineering Chemistry Research*, 41(10):2425–2430.

Bansal, L., Nelson, R., Yang, E., Jayaraman, A., and Hahn, J. (2013). Experimental design of systems involving multiple fluorescent protein reporters. *Chemical Engineering Science*, 101:191–198.

Bauer, I., Bock, H. G., Körkel, S., and Schlöder, J. (1999). Numerical methods for initial value problems and derivative generation for dae models with application to optimum experimental design of chemical processes. *Scientific Computing in Chemical Engineering II*, 2:282–289.

Bauer, I., Bock, H. G., Körkel, S., and Schlöder, J. P. (2000). Numerical methods for optimum experimental design in dae systems. *Journal of Computational and Applied Mathematics*, 120(1):1–25.

Ben-Zvi, A., McLellan, P. J., and McAuley, K. (2006). Identifiability of non-linear differential algebraic systems via a linearization approach. *The Canadian Journal of Chemical Engineering*, 84(5):590–596.

Berkholz, R., Röhlig, D., and Guthke, R. (2000). Data and knowledge based experimental design for fermentation process optimization. *Enzyme and Microbial Technology*, 27(10):784–788.

Biegler, L. T., Cervantes, A. M., and Wächter, A. (2002). Advances in simultaneous strategies for dynamic process optimization. *Chemical Engineering Science*, 57(4):575–593.

Box, G. E. P. and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, 9(1):57–71.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brockmann, D., Rosenwinkel, K.-H., and Morgenroth, E. (2008). Practical identifiability of biokinetic parameters of a model describing two-step nitrification in biofilms. *Biotechnology and Bioengineering*, 101(3):497–514.

Brown, M., He, F., and Wilkinson, S. J. (2010). Properties of the proximate parameter tuning regularization algorithm. *Bulletin of Mathematical Biology*, 72(3):697–718.

Brown, M., He, F., and Yeung, L. F. (2008). Robust measurement selection for biochemical pathway experimental design. *International Journal of Bioinformatics Research and Applications*, 4(4):400–416.

Brun, R., Reichert, P., and Künsch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 37(4):1015–1030.

Buzzi-Ferraris, G. and Forzatti, P. (1983). A new sequential experimental design procedure for discriminating among rival models. *Chemical Engineering Science*, 38(2):225–232.

Byrd, R. H., Hribar, M. E., and Nocedal, J. (1999). An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900.

Casey, F. P., Baird, D., Feng, Q., Gutenkunst, R. N., Waterfall, J. J., Myers, C. R., Brown, K. S., Cerione, R. A., and Sethna, J. P. (2007). Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. *IET Systems Biology*, 1(3):190–202.

Catania, F. and Paladino, O. (2009). Optimal sampling for the estimation of dispersion parameters in soil columns using an iterative genetic algorithm. *Environmental Modelling & Software*, 24(1):115–123.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.

Chianeh, H. A., Stigter, J., and Keesman, K. J. (2011). Optimal input design for parameter estimation in a single and double tank system through direct control of parametric output sensitivities. *Journal of Process Control*, 21(1):111–118.

Chiş, O., Banga, J. R., and Balsa-Canto, E. (2011). Genssi: a software toolbox for structural identifiability analysis of biological models. *Bioinformatics*, 27(18):2610–2611.

Chis, O.-T., Banga, J. R., and Balsa-Canto, E. (2011). Structural identifiability of systems biology models: a critical comparison of methods. *PloS one*, 6(11):e27755.

Cho, K.-H., Shin, S.-Y., Kolch, W., and Wolkenhauer, O. (2003). Experimental design in systems biology, based on parameter sensitivity analysis using a Monte Carlo method: A case study for the tnfα-mediated nf-κ b signal transduction pathway. *Simulation*, 79(12):726–739.

Chu, Y. and Hahn, J. (2007). Parameter set selection for estimation of nonlinear dynamic systems. *AIChE Journal*, 53(11):2858–2870.

Chu, Y. and Hahn, J. (2008). Parameter set selection via clustering of parameters into pairwise indistinguishable groups of parameters. *Industrial & Engineering Chemistry Research*, 48(13):6000–6009.

Chu, Y. and Hahn, J. (2012). Generalization of a parameter set selection procedure based on orthogonal projections and the d-optimality criterion. *AIChE Journal*, 58(7):2085–2096.

Chu, Y. and Hahn, J. (2013). Necessary condition for applying experimental design criteria to global sensitivity analysis results. *Computers & Chemical Engineering*, 48(Supplement C):280–292.

Chu, Y., Huang, Z., and Hahn, J. (2009). Improving prediction capabilities of complex dynamic models via parameter selection and estimation. *Chemical Engineering Science*, 64(19):4178–4185.

Chu, Y., Huang, Z., and Hahn, J. (2010). Global sensitivity analysis procedure accounting for effect of available experimental data. *Industrial & Engineering Chemistry Research*, 50(3):1294–1304.

Chung, S. H., Ma, D. L., and Braatz, R. D. (2000). Optimal model-based experimental design in batch crystallization. *Chemometrics and Intelligent Laboratory Systems*, 50(1):83–90.

Ciucci, F. (2013). Revisiting parameter identification in electrochemical impedance spectroscopy: Weighted least squares and optimal experimental design. *Electrochimica Acta*, 87:532–545.

Clyde, M. A. (2001). Experimental design: A Bayesian perspective. *International Encyclopedia Social and Behavioral Sciences*, 8(1):5075–5081.

Cukier, R. I., Levine, H. B., and Shuler, K. E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26(1):1–42.

de Brauwere, A., De Ridder, F., Gourgue, O., Lambrechts, J., Comblen, R., Pintelon, R., Passerat, J., Servais, P., Elskens, M., Baeyens, W., et al. (2009). Design of a sampling strategy to optimally calibrate a reactive transport model: Exploring the potential for escherichia coli in the scheldt estuary. *Environmental Modelling & Software*, 24(8):969–981.

Degenring, D., Froemel, C., Dikta, G., and Takors, R. (2004). Sensitivity analysis for the reduction of complex metabolism models. *Journal of Process Control*, 14(7):729–745.

Dejaegher, B. and Vander Heyden, Y. (2011). Experimental designs and their recent advances in set-up, data interpretation, and analytical applications. *Journal of Pharmaceutical and Biomedical Analysis*, 56(2):141–158.

Devore, J. L. (2015). *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

Di Maggio, J., Ricci, J. C. D., and Diaz, M. S. (2010). Parameter estimation in kinetic models for large scale metabolic networks with advanced mathematical programming techniques. *Computer Aided Chemical Engineering*, 28:355–360.

Dickinson, R. P. and Gelinas, R. J. (1976). Sensitivity analysis of ordinary differential equation systems—a direct method. *Journal of Computational Physics*, 21(2):123–143.

Dirion, J.-L., Reverte, C., and Cabassud, M. (2008). Kinetic parameter estimation from tga: Optimal design of tga experiments. *Chemical Engineering Research and Design*, 86(6):618–625.

Eghtesadi, Z. and McAuley, K. B. (2014). Mean square error based method for parameter ranking and selection to obtain accurate predictions at specified operating conditions. *Industrial & Engineering Chemistry Research*, 53(14):6033–6046.

Eghtesadi, Z. and McAuley, K. B. (2016). Mean-squared-error-based method for parameter ranking and selection with noninvertible fisher information matrix. *AIChE Journal*, 62(4):1112–1125.

Espie, D. and Macchietto, S. (1989). The optimal design of dynamic experiments. *AIChE Journal*, 35(2):223–229.

Esposito, W. R. and Floudas, C. A. (2000). Global optimization for the parameter estimation of differential-algebraic systems. *Industrial & Engineering Chemistry Research*, 39(5):1291–1310.

Faller, D., Klingmüller, U., and Timmer, J. (2003). Simulation methods for optimal experimental design in systems biology. *Simulation*, 79(12):717–725.

Farina, M., Findeisen, R., Bullinger, E., Bittanti, S., Allgower, F., and Wellstead, P. (2006). Results towards identifiability properties of biochemical reaction networks. In *Decision and Control, 2006 45th IEEE Conference on*, pages 2104–2109. IEEE.

Feng, X.-j. and Rabitz, H. (2004). Optimal identification of biochemical reaction networks. *Biophysical Journal*, 86(3):1270–1281.

Ferraris, G. B., Forzatti, P., Emig, G., and Hofmann, H. (1984). Sequential experimental design for model discrimination in the case of multiple responses. *Chemical Engineering Science*, 39(1):81–85.

Fiordalis, A. and Georgakis, C. (2013). Data-driven, using design of dynamic experiments, versus model-driven optimization of batch crystallization processes. *Journal of Process Control*, 23(2):179–188.

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, 1(3923):554.

Flaherty, P., Arkin, A., and Jordan, M. I. (2006). Robust design of biological experiments. In *Advances in Neural Information Processing Systems (NIPS)*, pages 363–370. MIT Press.

Fletcher, R. and Powell, M. J. (1963). A rapidly convergent descent method for minimization. *The computer journal*, 6(2):163–168.

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

Fricke, J., Pohlmann, K., Jonescheit, N. A., Ellert, A., Joksch, B., and Luttmann, R. (2013). Designing a fully automated multi-bioreactor plant for fast doe optimization of pharmaceutical protein production. *Biotechnology Journal*, 8(6):738–747.

Georgakis, C. (2013). Design of dynamic experiments: a data-driven methodology for the optimization of time-varying processes. *Industrial & Engineering Chemistry Research*, 52(35):12369–12382.

Gil, M. M., Miller, F. A., Silva, C. L., and Brandão, T. R. (2014). Application of optimal experimental design concept to improve the estimation of model parameters in microbial thermal inactivation kinetics. *Journal of Food Engineering*, 134:59–66.

Grewal, M. and Glover, K. (1976). Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions on Automatic Control*, 21(6):833–837.

Grewel, M. and Glover, K. (1976). Identifiability of linear and nonlinear dynamic sytems. *IEEE Transactions on Automatic Control*, 21(6):833–837.

Hagen, D. R., White, J. K., and Tidor, B. (2013). Convergence in parameters and predictions using computational experimental design. *Interface Focus*, 3(4):20130008.

He, F., Brown, M., and Yue, H. (2010). Maximin and Bayesian robust experimental design for measurement set selection in modelling biochemical regulatory systems. *International Journal of Robust and Nonlinear Control*, 20(9):1059–1078.

Hering, P. and Šimandl, M. (2010). Sequential optimal experiment design for neural networks using multiple linearization. *Neurocomputing*, 73(16):3284–3290.

Hibbert, D. B. (2012). Experimental design in chromatography: a tutorial review. *Journal of Chromatography B*, 910:2–13.

Holford, N. H. (2005). Fitting models to biological data using linear and non-linear regression: a practical guide to curve fitting. *Statistics in Medicine*, 24(17):2745–2746.

Holmberg, A. (1982). On the practical identifiability of microbial growth models incorporating Michaelis-Menten type nonlinearities. *Mathematical Biosciences*, 62(1):23–43.

Hosten, L. H. (1974). A sequential experimental design procedure for precise parameter estimation based upon the shape of the joint confidence region. *Chemical Engineering Science*, 29(11):2247 – 2252.

Huan, X. and Marzouk, Y. (2014). Gradient-based stochastic optimization methods in Bayesian experimental design. *International Journal for Uncertainty Quantification*, 4(6):479–510.

Huang, Y. and Wu, H. (2008). Bayesian experimental design for long-term longitudinal hiv dynamic studies. *Journal of Statistical Planning and Inference*, 138(1):105–113.

Huang, Z. J., Chu, Y., and Hahn, J. (2010). Model simplification procedure for signal transduction pathway models: An application to IL-6 signaling. *Chemical Engineering Science*, 65(6):1964–1975.

Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J., and Theis, F. (2013). High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. *Mathematical Biosciences*, 246(2):293–304.

Hunter, W. G. and Reiner, A. M. (1965). Designs for discriminating between two rival models. *Technometrics*, 7(3):307–323.

Iooss, B. and Lemaître, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems*, pages 101–122. Springer.

Jaqaman, K. and Danuser, G. (2006). Linking data to models: data regression. *Nature Reviews. Molecular Cell Biology*, 7(11):813.

Jia, J. and Yue, H. (2012). Measurement set selection of parameter estimation in biological system modelling-a case study of signal transduction pathways. *Journal-China University of Science and Technology*, 42(10):828–835.

Jin, Y., Yue, H., Brown, M., Liang, Y., and Kell, D. B. (2007). Improving data fitting of a signal transduction model by global sensitivity analysis. In *American Control Conference, 2007. ACC'07*, pages 2708–2713. IEEE.

Kennedy, J. (2011). Particle swarm optimization. In *Encyclopedia of machine learning*, pages 760–766. Springer.

Kolaczyk, E. D. and Csárdi, G. (2014). *Statistical analysis of network data with R*, volume 65. Springer.

Körkel*, S., Kostina, E., Bock, H. G., and Schlöder, J. P. (2004). Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes. *Optimization Methods and Software*, 19(3-4):327–338.

Kramer, A. and Radde, N. (2010). Towards experimental design using a Bayesian framework for parameter identification in dynamic intracellular network models. *Procedia Computer Science*, 1(1):1645–1653.

Kremling, A., Fischer, S., Gadkar, K., Doyle, F. J., Sauter, T., Bullinger, E., Allgöwer, F., and Gilles, E. D. (2004). A benchmark for methods in reverse engineering and model discrimination: Problem formulation and solutions. *Genome Research*, 14(9):1773–1785.

Kreutz, C. and Timmer, J. (2009). Systems biology: experimental design. *The FEBS Journal*, 276(4):923–942.

Kucherenko, S. (2005). Global sensitivity indices for nonlinear mathematical models. review. *Wilmott Mag*, 1:5661.

Kutalik, Z., Cho, K.-H., and Wolkenhauer, O. (2004). Optimal sampling time selection for parameter estimation in dynamic pathway modeling. *Biosystems*, 75(1):43–55.

Li, R., Henson, M. A., and Kurtz, M. J. (2004). Selection of model parameters for off-line parameter estimation. *IEEE Transactions on Control Systems Technology*, 12(3):402–412.

Lindner, P. F. O. and Hitzmann, B. (2006). Experimental design for optimal parameter estimation of an enzyme kinetic process based on the analysis of the fisher information matrix. *Journal of Theoretical Biology*, 238(1):111–123.

Ljung, L. (1998). System identification. In *Signal analysis and prediction*, pages 163–173. Springer.

Ljung, L. and Glad, T. (1994). On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265–276.

Ljung, L. and Wills, A. (2010). Issues in sampling and estimating continuous-time models with stochastic disturbances. *Automatica*, 46(5):925–931.

Lohmann, T., Bock, H. G., and Schloeder, J. P. (1992). Numerical methods for parameter estimation and optimal experiment design in chemical reaction systems. *Industrial & Engineering Chemistry Research*, 31(1):54–57.

Long, Q., Scavino, M., Tempone, R., and Wang, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259(Supplement C):24–39.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):231–242.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2008). Optimal designs for discriminating between some extensions of the Michaelis–Menten model. *Journal of Statistical Planning and Inference*, 138(12):3797–3804.

Massart, D. L., Kaufman, L., Rousseeuw, P. J., and Leroy, A. (1986). Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta*, 187:171–179.

McLean, K. A. and McAuley, K. B. (2012). Mathematical modelling of chemical processes—obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *The Canadian Journal of Chemical Engineering*, 90(2):351–366.

McLean, K. A., Wu, S., and McAuley, K. B. (2012). Mean-squared-error methods for selecting optimal parameter subsets for estimation. *Industrial & Engineering Chemistry Research*, 51(17):6105–6115.

McRae, G. J., Tilden, J. W., and Seinfeld, J. H. (1982). Global sensitivity analysis—a computational implementation of the fourier amplitude sensitivity test (fast). *Computers & Chemical Engineering*, 6(1):15–25.

Mintz, D. and Meer, P. (1991). Least median of squares regression. *Artificial Intelligence and Computer Vision*, 1(1):61.

Moles, C. G., Mendes, P., and Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, 13(11):2467–2474.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Motulsky, H. and Christopoulos, A. (2004). *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting*. Oxford University Press.

Murphy, E. F., Gilmour, S. G., and Crabbe, M. J. C. (2003). Efficient and accurate experimental design for enzyme kinetics: Bayesian studies reveal a systematic approach. *Journal of Biochemical and Biophysical Methods*, 55(2):155–178.

Omony, J., Mach-Aigner, A. R., de Graaff, L. H., van Straten, G., and van Boxtel, A. J. (2012). Evaluation of design strategies for time course experiments in genetic networks: case study of the XlnR regulon in aspergillus niger. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(5):1316–1325.

Otsu, T. (2008). Optimal experimental design criterion for discriminating semiparametric models. *Journal of Statistical Planning and Inference*, 138(12):4141–4150.

Pagendam, D. and Pollett, P. (2013). Optimal design of experimental epidemics. *Journal of Statistical Planning and Inference*, 143(3):563–572.

Paquet-Durand, O., Zettel, V., and Hitzmann, B. (2015). Optimal experimental design for parameter estimation of the peleg model. *Chemometrics and Intelligent Laboratory Systems*, 140:36–42.

Peifer, M. and Timmer, J. (2007). Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting. *IET Systems Biology*, 1(2):78–88.

Peleg, M., Yeh, I., and Altman, R. B. (2002). Modeling biological processes using workflow and petri net models. *Bioinformatics*, 18(6):825–837.

Petersen, B., Gernaey, K., and Vanrolleghem, P. A. (2001). Practical identifiability of model parameters by combined respirometric-titrimetric measurements. *Water Science and Technology*, 43(7):347–355.

Phair, R. D. (1997). Development of kinetic models in the nonlinear world of molecular cell biology. *Metabolism*, 46(12):1489–1495.

Pohjanpalo, H. (1978). System identifiability based on the power series expansion of the solution. *Mathematical Biosciences*, 41(1-2):21–33.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162.

Price, J., Hofmann, B., Silva, V. T., Nordblad, M., Woodley, J. M., and Huusom, J. K. (2014a). Mechanistic modeling of biodiesel production using a liquid lipase formulation. *Biotechnology Progress*, 30(6):1277–1290.

Price, J., Nordblad, M., Woodley, J. M., and Huusom, J. K. (2014b). Fed-batch feeding strategies for enzymatic biodiesel production. *IFAC Proceedings Volumes*, 47(3):6204–6209.

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

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929.

Reichert, P. and Vanrolleghem, P. (2001). Identifiability and uncertainty analysis of the river water quality model no. 1 (rwqm1). *Water Science and Technology*, 43(7):329–338.

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

Rojas, C. R., Welsh, J. S., Goodwin, G. C., and Feuer, A. (2007). Robust optimal experiment design for system identification. *Automatica*, 43(6):993–1008.

Ryan, E. G., Drovandi, C. C., Thompson, M. H., and Pettitt, A. N. (2014). Towards Bayesian experimental design for nonlinear models that require a large number of sampling times. *Computational Statistics & Data Analysis*, 70(Supplement C):45–60.

Saltelli, A. (2004). Global sensitivity analysis: an introduction. In *Proc. 4th International Conference on Sensitivity Analysis of Model Output (SAMO'04)*, pages 27–43.

Saltelli, A. and Bolado, R. (1998). An alternative way to compute fourier amplitude sensitivity test (fast). *Computational Statistics & Data Analysis*, 26(4):445–460.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

Schittkowski, K. (2007). Experimental design tools for ordinary and algebraic differential equations. *Industrial & Engineering Chemistry Research*, 46(26):9137–9147.

Schwaab, M., Silva, F. M., Queipo, C. A., Barreto, A. G., Nele, M., and Pinto, J. C. (2006). A new approach for sequential experimental design for model discrimination. *Chemical Engineering Science*, 61(17):5791–5806.

Seagren, E., Kim, H., and Smets, B. (2003). Identifiability and retrievability of unique parameters describing intrinsic andrews kinetics. *Applied Microbiology and Biotechnology*, 61(4):314–322.

Silvey, S. (2013). *Optimal design: an introduction to the theory for parameter estimation*, volume 1. Springer Science & Business Media.

Sjögren, E., Nyberg, J., Magnusson, M. O., Lennernäs, H., Hooker, A., and Bredberg, U. (2011). Optimal experimental design for assessment of enzyme kinetics in a drug discovery screening environment. *Drug Metabolism and Disposition*, 39(5):858–863.

Skanda, D. and Lebiedz, D. (2010). An optimal experimental design approach to model discrimination in dynamic biochemical systems. *Bioinformatics*, 26(7):939–945.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280.

Storn, R. and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359.

Strigul, N., Dette, H., and Melas, V. B. (2009). A practical guide for optimal designs of experiments in the monod model. *Environmental Modelling & Software*, 24(9):1019–1026.

Sturm, J. F. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1-4):625–653.

Sun, J., Garibaldi, J. M., and Hodgman, C. (2012). Parameter estimation using metaheuristics in systems biology: a comprehensive review. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(1):185–202.

Sun, N.-Z. and Sun, A. (2015). *Model Calibration and Parameter Estimation: For Environmental and Water Resource Systems*. Springer.

Thompson, D. E., McAuley, K. B., and McLellan, P. J. (2009). Parameter estimation in a simplified mwd model for hdpe produced by a ziegler-natta catalyst. *Macromolecular Reaction Engineering*, 3(4):160–177.

Tommasi, C. (2009). Optimal designs for both model discrimination and parameter estimation. *Journal of Statistical Planning and Inference*, 139(12):4123–4132.

Tommasi, C. and López-Fidalgo, J. (2010). Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics & Data Analysis*, 54(1):143–150.

Tulsyan, A., Forbes, J. F., and Huang, B. (2012). Designing priors for robust Bayesian optimal experimental design. *Journal of Process Control*, 22(2):450–462.

Turányi, T. (1990). Sensitivity analysis of complex kinetic systems. tools and applications. *Journal of Mathematical Chemistry*, 5(3):203–248.

Ucinski, D. and Bogacka, B. (2004). Heteroscedastic t-optimum designs for multiresponse dynamic models. In *Proceedings of the 7th International Workshop on Model-Oriented Design and Analysis, Physica Verlag Germany, Heidelberg*, pages 191–199. Physica Verlag, Heidelberg, Germany.

Vajda, S., Godfrey, K. R., and Rabitz, H. (1989). Similarity transformation approach to identifiability analysis of nonlinear compartmental models. *Mathematical Biosciences*, 93(2):217–248.

Vajda, S. and Rabitz, H. (1989). State isomorphism approach to global identifiability of nonlinear systems. *IEEE Transactions on Automatic Control*, 34(2):220–223.

Van Derlinden, E., Mertens, L., and Van Impe, J. F. (2013). The impact of experiment design on the parameter estimation of cardinal parameter models in predictive microbiology. *Food Control*, 29(2):300–308.

van Riel, N. A. (2006a). Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Briefings in Bioinformatics*, 7(4):364–374.

van Riel, N. A. (2006b). Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Briefings in bioinformatics*, 7(4):364–374.

Vanlier, J., Tiemann, C., Hilbers, P., and Van Riel, N. (2013). Parameter uncertainty in biochemical models described by ordinary differential equations. *Mathematical Biosciences*, 246(2):305–314.

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J., and van Riel, N. A. W. (2012). A Bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142.

Vassiliadis, V., Sargent, R., and Pantelides, C. (1994a). Solution of a class of multistage dynamic optimization problems. 1. problems without path constraints. *Industrial & Engineering Chemistry Research*, 33(9):2111–2122.

Vassiliadis, V., Sargent, R., and Pantelides, C. (1994b). Solution of a class of multistage dynamic optimization problems. 2. problems with path constraints. *Industrial & Engineering Chemistry Research*, 33(9):2123–2133.

Versyck, K. J. and Van Impe, J. F. (1998). Trade-offs in design of fed-batch experiments for optimal estimation of biokinetic parameters. In *Control Applications, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 51–55. IEEE.

Villaverde, A. F., Henriques, D., Smallbone, K., Bongard, S., Schmid, J., Cicin-Sain, D., Crombach, A., Saez-Rodriguez, J., Mauch, K., Balsa-Canto, E., et al. (2014). Biopredynbench: benchmark problems for kinetic modelling in systems biology. *arXiv preprint arXiv:1407.5856*.

Walter, E. and Lecourtier, Y. (1982). Global approaches to identifiability testing for linear and nonlinear state space models. *Mathematics and Computers in Simulation*, 24(6):472–482.

Walter, É. and Pronzato, L. (1990). Qualitative and quantitative experiment design for phenomenological models—a survey. *Automatica*, 26(2):195–213.

Walter, E. and Pronzato, L. (1996). On the identifiability and distinguishability of nonlinear parametric models. *Mathematics and Computers in Simulation*, 42(2-3):125–134.

Weijers, S. R. and Vanrolleghem, P. A. (1997). A procedure for selecting best identifiable parameters in calibrating activated sludge model no. 1 to full-scale plant data. *Water Science and Technology*, 36(5):69–79.

Wiens, D. P. (2009). Robust discrimination designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):805–829.

Wu, H., Zhu, H., Miao, H., and Perelson, A. S. (2008). Parameter identifiability and estimation of hiv/aids dynamic models. *Bulletin of Mathematical Biology*, 70(3):785–799.

Wu, S., McAuley, K., and Harris, T. (2011a). Selection of simplified models: I. analysis of model-selection criteria using mean-squared error. *The Canadian Journal of Chemical Engineering*, 89(1):148–158.

Wu, S., McAuley, K., and Harris, T. (2011b). Selection of simplified models: Ii. development of a model selection criterion based on mean squared error. *The Canadian Journal of Chemical Engineering*, 89(2):325–336.

Yao, K. Z., Shaw, B. M., Kou, B., McAuley, K. B., and Bacon, D. (2003). Modeling ethylene/butene copolymerization with multi-site catalysts: parameter estimability and experimental design. *Polymer Reaction Engineering*, 11(3):563–588.

Yen, J., Liao, J. C., Lee, B., and Randolph, D. (1998). A hybrid approach to modeling metabolic systems using a genetic algorithm and simplex method. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(2):173–191.

Yu, H., Yue, H., and Halling, P. (2015). Optimal experimental design for an enzymatic biodiesel production system. *IFAC-PapersOnLine*, 48(8):1258–1263.

Yue, H., Brown, M., Knowles, J., Wang, H., Broomhead, D. S., and Kell, D. B. (2006). Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: A case study of an NF-κB signalling pathway. *Molecular BioSystems*, 2(12):640–649.

Yue, H., Halling, P., and Yu, H. (2013). Model development and optimal experimental design of a kinetically controlled synthesis system. *IFAC Proceedings Volumes*, 46(31):327–332.

Zak, D. E., Gonye, G. E., Schwaber, J. S., and Doyle, F. J. (2003). Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. *Genome Research*, 13(11):2396–2405.

Zen, M.-M. and Tsai, M.-H. (2004). Criterion-robust optimal designs for model discrimination and parameter estimation in fourier regression models. *Journal of Statistical Planning and Inference*, 124(2):475–487.

Zhang, T., Zhang, D., Li, Z., and Cai, Q. (2010). Evaluating the structural identifiability of the parameters of the EBPR sub-model in ASM2d by the differential algebra method. *Water Research*, 44(9):2815–2822.

Zhu, Z. and Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1):24–44.

Zi, Z. (2011). Sensitivity analysis approaches applied to systems biology models. *IET Systems Biology*, 5(6):336–346.

Zou, R. and Ghosh, A. (2006). Automated sensitivity analysis of stiff biochemical systems using a fourth-order adaptive step size rosenbrock integration method. *IEE Proceedings-Systems Biology*, 153(2):79–90.

Zullo, L. (1991). *Computer Aided Design of Experiments: An Engineering Approach*. PhD thesis, Imperial College London.

Appendix A

Identifiability analysis methods

Structural identifiability analysis

Generating series approach

Consider a general dynamic model

$$\begin{aligned}\dot{\mathbf{X}}(t) &= \mathbf{f}(\mathbf{X}(t),\boldsymbol{\theta}) + \sum_{i=1}^{n_u} g_i(\mathbf{X}(t),\boldsymbol{\theta})\,u_i(t)\\ \mathbf{Y} &= \mathbf{h}(\mathbf{X}(t),\boldsymbol{\theta}) + \boldsymbol{\xi}(t)\end{aligned}$$

where g_i and $\mathbf{h}$ are analytic on the manifold over which $\mathbf{X}(t)$ evolves, so that the model output can be expanded in series with respect to time and inputs. The coefficients of this series are $\mathbf{h}(\mathbf{X}(0),\boldsymbol{\theta})$ and its Lie derivatives

$$L_{f_{j0}} \cdots L_{f_{jk}}\mathbf{h}(\mathbf{X}(t_0),\boldsymbol{\theta})$$

where

$$L_{\mathbf{f}}\mathbf{h}\left(\mathbf{X}\left(t_0\right),\boldsymbol{\theta}\right)=\sum_{j=1}^{n} g_j\left(\mathbf{X}\left(t_0\right),\boldsymbol{\theta}\right)\cdot\frac{\partial}{\partial x_j}\mathbf{h}\left(\mathbf{X}\left(t_0\right),\boldsymbol{\theta}\right)$$

in which g_i is the j-th component of $\mathbf{g}$. Let $s(\boldsymbol{\theta})$ be the vector of all coefficients of the series. One can therefore test the identifiability by calculating the number of solutions for $\boldsymbol{\theta}$.

Practical identifiability analysis

Orthogonal sensitivity selection algorithm

The orthogonalisation based parameter subset selection method is based on the measure of orthogonal parameter sensitivities. In other words, the parameter pair correlations have been removed from original local sensitivity matrix and measurement is focused on the independent effect of parameters to system outputs. The numerical procedure of this method is given here. Firstly, a two-dimensional ($[n \times N,\ p]$) normalized parameter sensitivity matrix $\bar{\mathbf{S}}$ is obtained from the solution of equation 2.12 and 2.16, Each column of which represents one parameter.

Then the sensitivity effect of each parameter can be determined by the magnitude (norm, equation 2.19) of each column of matrix $\bar{\mathbf{S}}$. As the parameters are correlated with each other, removing one parameter changes the effect of remaining parameters. Therefore, the orthogonalisation based forward selection method is a useful tool to identify important parameters as described below:

1. Using Euclidean norm (equation 2.19) to calculate the magnitude of each column of $\bar{\mathbf{S}}$. The parameter corresponding to the column with maximum magnitude is the first identifiable parameter. Set $k = 1$.

2. Put the k columns from $\bar{\mathbf{S}}$ that correspond to parameters that have been identified into matrix $\mathbf{X}_k$.

3. Use $\mathbf{X}_k$ to calculate the ordinary least-square prediction of matrix $\bar{\mathbf{S}}$:

$$\hat{\bar{\mathbf{S}}}_k = \mathbf{X}_k \left(\mathbf{X}_k^T \mathbf{X}_k\right)^{-1} \mathbf{X}_k \bar{\mathbf{S}}$$

   and calculate the residual matrix

$$R_k = \bar{\mathbf{S}} - \hat{\bar{\mathbf{S}}}_k$$

4. Calculate the magnitude of each column in R_k. The *(k+1)*-th most identifiable parameter corresponds to the column in R_k with the largest magnitude.

5. Increase k by 1 and add the column of $\bar{\mathbf{S}}$ that correspond to the *(k+1)*-th parameter to matrix $\mathbf{X}_k$.

6. Repeat step 3-5 for all parameters or until the maximum magnitude in $\mathbf{X}_k$ is less than a predefined threshold.

This procedure selects model parameter sequentially, which is also called forward subset selection method. One closely connected method is called backward elimination method. Rather than selecting the most important parameter in a 1-step fashion and putting it into selected matrix $\mathbf{X}_k$, this method is to remove the most unimportant parameter one at a time and initially $\mathbf{X}_k$ contains all parameter sensitivity columns.

Backwards elimination procedure

1. Start with $\mathbf{X}_k = \bar{\mathbf{S}}$, where all parameters are included in $\mathbf{X}_k$ and the Hessian matrix is $\mathbf{H}_k = \mathbf{X}_k^T \mathbf{X}_k$.

2. Update the Hessian matrix $\mathbf{H}_k$ with j-th parameter be removed:

$$\mathbf{H}_{k-1}^{-1} = [\mathbf{H}_k^{-1}] - \frac{[\mathbf{H}_k^{-1}]_j [\mathbf{X}_k^T]_j x_j x_j^T [\mathbf{X}_k]_j [\mathbf{H}_k^{-1}]_j}{x_j^T \left(\mathbf{I} - [\mathbf{X}_k]_j [\mathbf{H}_k^{-1}][\mathbf{H}_k^{-1}]_j [\mathbf{X}_k^T]_j\right) x_j}$$

   where $[\cdot]_j$ denotes the sub-matrix with j-th parameter removed.

3. Remove the selected j-th (index j^*) parameter with the smallest t-ratio in magnitude:

$$j^* = \arg\min_j |\frac{\theta_j}{\sigma(\theta_j)}| = \arg\min_j |\frac{1}{\sigma_n^2 \sqrt{[\mathbf{H}_{k-1}^{-1}]_{j,j}}}|$$

4. Repeat step 2-3 until the effect of removed parameter is significant. (bigger than predefined t-ratio threshold).

Collinearity index algorithm

The numerical procedure of collinearity index method to select identifiable parameter subset is given as follows:

1. The normalised parameter sensitivity matrices obtain from equation 2.12 is further rescaled as:

$$\tilde{\mathbf{S}} = \begin{bmatrix} \tilde{\mathbf{S}}(1) \\ \tilde{\mathbf{S}}(2) \\ \vdots \\ \tilde{\mathbf{S}}(N) \end{bmatrix}, \quad \tilde{\mathbf{S}}(k) = \begin{bmatrix} \tilde{s}_{1,1}(k) & \tilde{s}_{1,2}(k) & \cdots & \tilde{s}_{1,p}(k) \\ \tilde{s}_{2,1}(k) & \tilde{s}_{2,2}(k) & \cdots & \tilde{s}_{2,p}(k) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{s}_{n,1}(k) & \tilde{s}_{n,2}(k) & \cdots & \tilde{s}_{n,p}(k) \end{bmatrix}, \quad \tilde{s}_{i,j} = \frac{\bar{s}_{i,j}}{\sqrt{\sum_{k=1}^{N} \bar{s}_{k,j}^2}}$$

   where $\tilde{\mathbf{S}}$ is called non-dimensional parameter sensitivities.

2. For all possible parameter subset combinations $K \leq p$, the non-dimensional FIM is calculated and the corresponding collinearity index can be determined.

$$\Lambda = \tilde{\mathbf{S}}(K)^T \tilde{\mathbf{S}}(K)$$
$$\gamma(K) = \frac{1}{\sqrt{\lambda_{min}(\Lambda)}}$$

3. A change in the output vector caused by a shift of a parameter can be compensated up to a fraction of 1 divided by the collinearity index $\gamma(K)$ by appropriate changes in the other parameter in K. Therefore, a high value of $\gamma(K)$ indicates parameter subset K are poorly identifiable. With given threshold for the collinearity index, all possible identifiable parameter subsets can be determined.

Principle component analysis

The identifiable parameter subset selection based on principle component analysis is briefly described as follows:

The non-dimensional sensitivity matrix $\bar{\mathbf{S}}$ is calculated from equation 2.12 and the non-dimensional FIM can be determined as $\bar{\mathbf{S}}^T \bar{\mathbf{S}}$.

The eigenvalues and eigenvectors are obtained for $\bar{\mathbf{S}}^T \bar{\mathbf{S}}$, and the eigenvalues are ordered from smallest to largest:

$$|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_p|$$

The corresponding eigenvectors are columns in the matrix:

$$\Gamma = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p) = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,p} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p,1} & v_{p,2} & \cdots & v_{p,p} \end{bmatrix}$$

The parameter subset can be selected by using three different methods:

- Starting from the smallest eigenvalue, the parameters are ranked from least estimable to most estimable by selecting the parameter with the largest weight in the corresponding eigenvector. In this way, the least estimable parameter is selected, followed by the second least estimable parameter and so on, until all of the parameters are ranked.

- Parameters are ranked from least estimable to most estimable by comparing row-wise sums of squares from Γ. The first parameter selected is the one with the largest squared value in the first eigenvector. The j-th least estimable parameter is selected by comparing the row-wise sums of squares of first j elements in each row of Γ.

- Starting from the eigenvector that corresponds to the largest eigenvalue, the most estimable parameter is selected by finding the largest absolute value in the corresponding eigenvector. The second most estimable parameter is selected by checking eigenvector that corresponds to the second largest eigenvalue and so on.

Appendix B

Preliminary analysis on multi-objective OED

Integration of OED and process optimisation

The OED techniques we have discussed so far are focused on maximising data information through the design of experimental manipulations and measurement strategies. The ultimate purpose is to develop a perfect model that we expect to cover all the important features of the process in all kinds of circumstances. The biochemical model developed by following the OED procedures could present well the real bioprocess under various experimental conditions. However, from the perspective of experimentalists, they would be more concerned about how well the model can describe the process in a particular operation condition, e.g. the ability of the model to predict the conditions maximising the information of a product of interest. To this purpose, the experimental design should be combined with process optimisation so that one can develop a model that can be more specific for the description of the process in possibly the real operation condition.

In this work, a new sequential procedure is introduced to combine experimental design with process optimisation in order to achieve both reducing parameter estimation errors and maximising the production of variables of interest. As mentioned in Section 6.2, the experimental manipulations such as initial input conditions and external input variables will change the behaviour of system dynamics, but the change of measurement strategy will not change the performance of the process. Therefore, in the input design we can mainly focus on the optimisation of process performance, from which the optimal input values can be obtained which would be close to the condition when the actual experimentation is performed. In the second step, the observation design is conducted in order to choose the optimal measurement strategy based on FIM so that the generated data from designed experiment could possibly reduce parameter estimation errors. This sequential design process is iterated until satisfactory results are obtained.

The main advantage of this method is that the real point at which the experiments are performed is fairly close to optimal initial conditions which can lead to the best performance of the process and at the same time observation design will make sure informative data can be obtained to increase parameter estimation precision. The final calibrated model could possibly represent the real process at the optimal condition more precisely than when the process optimisation is not considered.

The proposed method is applied to the enzyme reaction system which is described in Section 4.1. In the first step, the input design is mainly focused on the maximisation of production of Q where the time to reach that optimal point is constrained at 1000 seconds in order to make sure that the enzymatic process is not too slow, the objective can be described as:

$$\mathbf{X}_0^* = \arg\max_{\mathbf{X}_0 \in \mathbf{\Phi}} Q_{max} \qquad s.t. \quad t_{max} \leq 1000 \tag{B.1}$$

where $\mathbf{X}_0$ is the vector of input variables. In this enzyme reaction system, three initial input states (S_0, N_0, E_0) will be considered. The objective is constrained by model DAEs (4.3). Then in the second step, the observation design described in equation (6.2) is applied in order to find the best 50 measurement time points and valuable measurement set. The D-optimal criterion is applied in this case. The simulation result is shown in the following table. It can be found that by using our method the optimal sampling time points are close to those from integrated design result which are shown in Table 6.3 and the most valuable measurement states are S and Q. The predicted parameter estimation errors for k_{-3} and k_5W can be reduced to less than 5%. Also, the maximum value of Q is 0.62 $mol \cdot L^{-1}$ under the designed input condition.

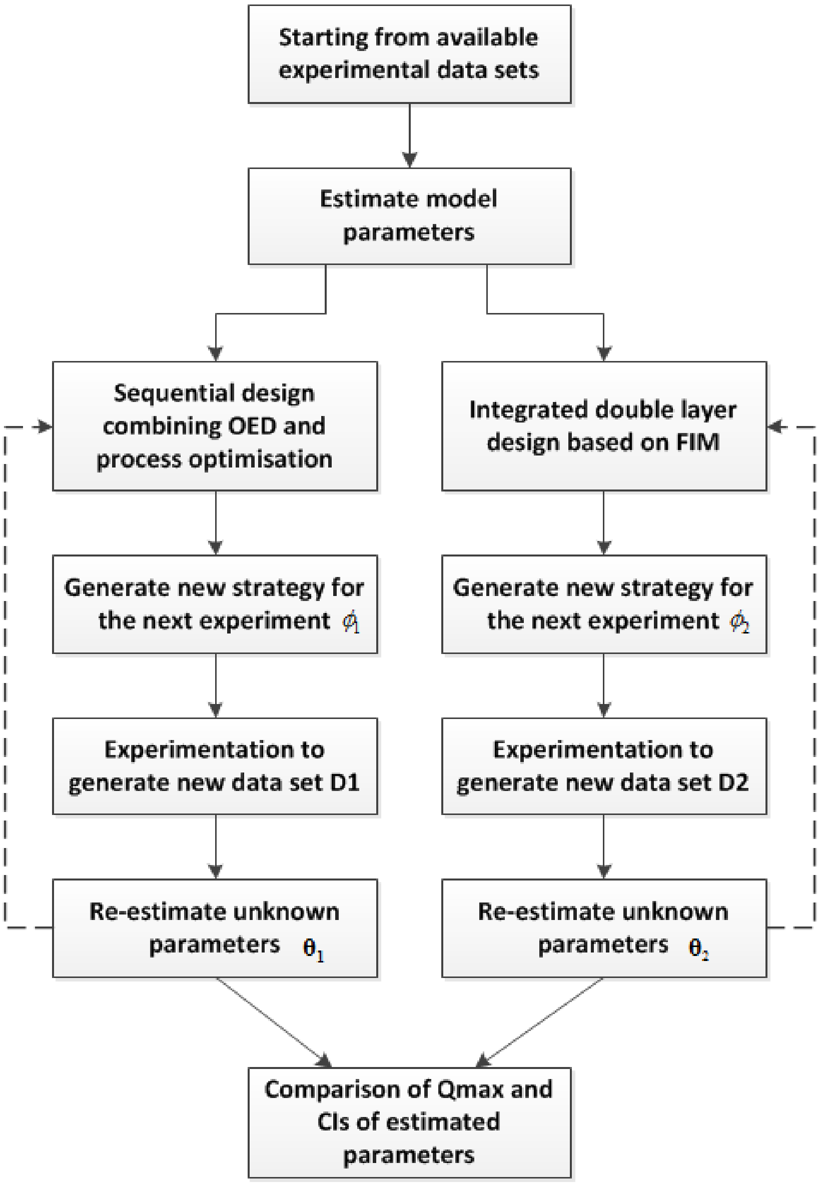
Optimal experimental condition with combination of observation design and process optimisation for the enzyme reaction model

<table>
<tr><th>Numerical Strategies</th><th>Input condition ($mol \cdot L^{-1}$)</th><th>Measurement set</th><th colspan="2">Sampling time points (sec)</th></tr>
<tr><td rowspan="2">OED combined with process optimisation</td><td rowspan="2">S_0: 1
E_0: 1.53e-5
N_0: 1</td><td>S</td><td colspan="2">630:30:990, 1020, 1050, 4440:30:4920</td></tr>
<tr><td>Q</td><td colspan="2">5490:30:6000</td></tr>
<tr><td>D-value</td><td colspan="4">2.24e-8</td></tr>
<tr><td>Q_{max}</td><td colspan="4">0.62 $mol \cdot L^{-1}$</td></tr>
<tr><td colspan="5"><b>Predicted parameter estimation errors</b> (%)</td></tr>
<tr><td>k_2
12%</td><td colspan="2">k_{-3}
3.63%</td><td colspan="2">k_5W
2.7%</td></tr>
</table>

It should be emphasised that the proposed method does not need to set any weighting factors from experience to balance between experimental design for parameter estimation and process optimisation. This multi-objective design problem is solved separately yet combined together in a sequential process. The input design which will change system dynamics is mainly focused on the optimisation of process performance and then the observation design is aimed at selecting the most informative data in order to facilitate parameter estimation. This input and observation design is iterated until the final calibrated model meets the requirement. On the one hand experiments are always performed around the optimal input condition that will possibly lead to the best process performance; on the other hand valuable data can be obtained from observation design. In this way, the final model can well describe the real process at the optimal operation point.

In order to compare the proposed sequential strategy, which combines OED for parameter estimation and process optimisation, to the integrated double layer OED method, which is purely based on the optimisation of a scalar function of FIM, the two approaches are implemented to the enzyme reaction system starting from the same parameter values and simulated data sets. The detailed procedure is shown in the following figure. The a priori experimental data are generated from three different input conditions, which are $[S_0 = 1mol \cdot L^{-1}, N_0 = 1mol \cdot L^{-1}]$, $[S_0 = 1mol \cdot L^{-1}, N_0 = 0.5mol \cdot L^{-1}]$, and $[S_0 = 0.5mol \cdot L^{-1}$, $N_0 = 1mol \cdot L^{-1}]$ respectively. In all three experiments, the input value of enzyme is $E_0 = 1.5e-5mol \cdot L^{-1}$; Q is selected to be the only measurement set and the sampling time strategy is $[T_{sp} = 0, 15, 30, 60, 120, 240, 480, 1000, 2000, 4000](sec)$. 5% relative measurement errors and 0.0001 $mol \cdot L^{-1}$ absolute errors are added to the simulated data.

To simplify the calculation and avoid the ill-posed parameter estimation problem, we assume that only three parameters $[k_2, k_{-3}, k_5W]$ are unknown. Based on the available experimental data, the LSE by using the interior-point optimisation method is implemented which gives these unknown parameter values as: $k_2 = 103.46$, $k_{-3} = 137.72$, and $k_5W = 6.62e3$.


Starting from available experimental data sets
Estimate model parameters
Sequential design combining OED and process optimisation
Integrated double layer design based on FIM
Generate new strategy for the next experiment ϕ_1
Generate new strategy for the next experiment ϕ_2
Experimentation to generate new data set D1
Experimentation to generate new data set D2
Re-estimate unknown parameters θ_1
Re-estimate unknown parameters θ_2
Comparison of Qmax and CIs of estimated parameters

Numerical procedure for comparison of two OED strategies

Then two different experimental design methods are applied to obtain the best experimental strategy for the next experiment. It is expected to obtain the best twenty sampling time points in total for all five measurable state variables (S, P, N, Q and R) and the best initial input values for S_0, N_0 and E_0. In this simulation, the interior-point method is employed for the parameter estimation; the PSA algorithm is applied for the calculation of input values and the observation design is solved for by the Powell's method. The simulation result is given in the following table.

Optimal experimental conditions by using two different OED strategies

OED Strategies	Input condition ($mol \cdot L^{-1}$)	Measurement set	Sampling time points (sec)	D-values
Integrated design	$S_0 = 1$	S	5460:30:5610	1.02e-4
	$N_0 = 0.05$	N	960:30:1140	
	$E_0 = 4.15e-6$	Q	5850:30:6000	
OED combined with process optimisation	$S_0 = 1$	S	780:30:930, 2730:30:2910	1.77e-5
	$N_0 = 1$			
	$E_0 = 1.55e-5$	Q	5430:30:5610	

It is obvious that these two design methods indicate totally different initial input conditions and measurements for the next experiment. Small values of N_0 and E_0 are suggested from the integrated design, while the sequential design shows that high value of N_0 is preferred. Also, in terms of the exact observation strategy three state variables, S, N and Q, are selected to be measured from integrated design while the sequential design indicates that only the measurement of S and Q is necessary. In addition, the exact sampling time points for each state variable are also very different between these two design methods. Based on the D-optimal values, it is expected that integrated design could theoretically provide more precise parameter estimates with additional data from next experiment.

To compare the performance of parameter estimation as well as maximal value of Q, another two data sets are generated based on the above two design results, with the same measurement errors as described before. The unknown parameters are re-estimated with additional data sets and then the optimal value of Q is obtained from PSA optimisation, which are shown as follows:

comparison of OED results between two OED strategies

OED Strategies	True parameters and Q_{max}	Estimated parameters	Uncertainties	$Q_{max}(mol \cdot L^{-1})$
Integrated design	$k_2 = 100$, $k_{-3} = 200$, $k_5W = 5e3$, $Q_{max} = 0.624$	$k_2 = 101.02$, $k_{-3} = 210.78$, $k_5W = 4.94e3$	$k_2 : 1.02\%$, $k_{-3} : 10.78\%$, $k_5W : 1.12\%$	0.588
OED combined with process optimisation		$k_2 = 100.91$, $k_{-3} = 189.54$, $k_5W = 5.32e3$	$k_2 : 0.91\%$, $k_{-3} : 10.46\%$, $k_5W : 7.28\%$	0.624

It can be seen that with the re-estimated parameter values, the maximal production of Q is closer to the true optimum of Q using the combined design strategy. By looking at the parameter estimation uncertainties, the uncertainties of k_2 and k_{-3} from both designs are close to each other, but k_5W contains larger uncertainty from the combined design than that from the integrated design. The integrated design strategy can generate more data information for parameter estimation than the combined design strategy. However, the OED combined with process optimisation provides another possibility to design the input. Based on the experimenters' experience, if the uncertainties of estimated model parameters is satisfied, the strategy of OED combined with process optimisation might be useful as it can provide additional information on the product yield of interest.

Appendix C

Preliminary analysis on robust experimental design

Global sensitivity analysis

An important aspect of OED for parameter estimation that has not been focused on so far in this thesis is the dependency of the design on the model parameters. Indeed, the basis for optimal experimental design is the Fisher Information Matrix (FIM) which is calculated from local sensitivity functions, the partial derivatives of the model outputs to the parameters. For non-linear models, these partial derivatives are still functions of the model parameters which means that the FIM is directly influenced by the parameter values themselves. Therefore, all experimental designs based on FIM properties are called local designs. The effectiveness of the design thus depends on how close the model parameters are to those of the real system. Furthermore, important parameters selected based on local sensitivity analysis directly depend on the choice of parameter values. When model parameters are uncertain and contain large uncertainties, it is necessary to analyse parameter effects on model outputs along the whole parameter range rather than at a guessed parameter value. Therefore, global sensitivity analysis (GSA) is required to comprehensively investigate parameter effects on model outputs of interest. In addition, previous work rarely discussed the connection of global sensitivity analysis and optimal experimental design. Global sensitivity analysis is mostly treated as a pre-screening step to select important parameters. Robust experimental design with the consideration of parameter uncertainties will be investigated. Firstly, two global sensitivity algorithms will be applied and compared in the biochemical context for the selection of important parameters. Then several robust experimental design methods will be applied and compared in the design of the sampling strategy for the enzyme reaction system. In particular, the integration of global sensitivity algorithm and FIM based experimental design is discussed.

Global sensitivity approaches are alternatives to local sensitivity analysis to quantify parameter effects on the model outputs. The main advantage of GSA over LSA from the experimental design point of view is that parameters can be varied simultaneously over their entire uncertainty range to investigate their effects on the outputs. Several GSA methods have been developed in the last few decades. GSA aims at apportioning the output uncertainty to the uncertainty in the model parameter values. A global sensitivity technique has the advantage of incorporating the influence of the whole variation range of parameters while providing quantitative results, which indicate the significance of each parameter. The effect on the outputs of changing one parameter while all the others are varied as well is evaluated by GSA and this can help in discovering parameter interactions in a model. In this work, two global sensitivity analysis methods are employed which are the Morris screening method and Sobol's method.

The Morris screening method (Morris, 1991) is a measurement tool of global sensitivity which is based on the so called elementary effect (EE). A number of values of EEs for each parameter can be obtained through a predefined randomly selected sampling strategy. The distribution of EEs from i-th parameter is denoted as F_i. The sensitivities are measured in two parts: μ_i, the mean of EEs is an estimate of the overall effect of the i-th parameter on model outputs; σ_i, the standard deviation of EEs is an evaluation of the ensemble of influences of the i-th parameter, which is attributed to the interactions with other parameters. These two measures will be used to indicate which parameters should be considered important.

Consider a general model with p parameters and an output y, $y = f(\theta_1, \theta_2, \cdots \theta_p)$. Each parameter has an uncertainty region which is scaled from 0 to 1; and it may take values from $0, 1/(p-1), 2/(p-1), \cdots, 1$. Then the EE of i-th parameter is defined as

$$EE_i(\boldsymbol{\theta}) = \frac{f(\theta_1, \cdots, \theta_i + \Delta, \cdots, \theta_p) - f(\boldsymbol{\theta})}{\Delta} \tag{C.1}$$

where Δ is a predetermined multiple of $1/(p-1)$ and is taken to be $\Delta = p/(2p-2)$ (Morris, 1991). Producing a value for F_i requires random selection of a value of each θ_i from the grid and evaluation of y twice, one at the selected parameter values, the other after increasing θ_i by Δ. The difference between these two runs yields one elementary effect. The calculation will be repeated γ times to produce a random sample of γ EEs for F_i. Owing to its designed sampling strategies, this method is computationally cheap as it requires a relatively small number of model evaluations.

Sobol's method (Sobol, 2001) is a popular sensitivity measure based on analysis of variance (ANOVA). In general, variance-based sensitivity analysis methods aim to quantify the amount of variance that each parameter contributes to the unconditional variance of model outputs. For Sobol's sensitivity analysis method, the variances caused by either a single parameter or by the interactions of two or more parameters are expressed as sensitivity indices. For a model represented by equation (2.1) with p parameters (for brevity we ignore time and state variables and the model is expressed as $f(\boldsymbol{\theta})$), Sobol suggested the decomposition of the model function into summands of increasing dimensionality:

$$f(\boldsymbol{\theta}) = f_0 + \sum_{i=1}^{p} f_i(\theta_i) + \sum_{i=1}^{p} \sum_{j=i+1}^{p} f_{i,j}(\theta_i, \theta_j) + \cdots + f_{1,2,\cdots,p}(\theta_1, \cdots, \theta_p) \tag{C.2}$$

where f_0 equals to the expectation value of the output. The total unconditional variance can then be determined as:

$$V = \int_{\Omega^p} f^2(\boldsymbol{\theta}) d\boldsymbol{\theta} - f_0^2 \tag{C.3}$$

with Ω^p the p-dimensional hypercube space of model parameters. The partial variances, which are the components of the total variance decomposition, are computed from each of the terms in equation (C.2) as

$$V_{i_1,\cdots,i_k} = \int_{i_1} \cdots \int_{i_k} f_{i_1,\cdots,i_k}(\theta_{i_1}, \cdots, \theta_{i_k}) d\theta_{i_1} \cdots d\theta_{i_k} \tag{C.4}$$

where $1 \leq i_1 \leq \cdots \leq i_k \leq p$. With the assumption that parameters are mutually orthogonal, the variance of outputs to parameters can be decomposed as:

$$V = \sum_{i=1}^{p} V_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} V_{i,j} + \cdots + V_{1,2,\cdots,p} \tag{C.5}$$

In this way, the variance contributions to the total output variance of individual parameters and parameter interactions can be determined. These contributions are characterized by the ratio of the partial variance to the total variance, the Sobol sensitivity indices:

First order SI: $S_i = \frac{V_i}{V}$

Second order SI: $S_{i,j} = \frac{V_{i,j}}{V}$

Total order SI: $S_{T_i} = S_i + \sum_{j \neq i} S_{i,j} + \cdots$

The first order index is a measure for the variance contribution of the individual parameter θ_i to the total model variance which is also called the 'main effect'. S_{T_i} is the result of the main effect of θ_i and all its interactions with other parameters. The Monte Carlo sampling is applied in this research to approximate the integral as it is impossible to calculate the variance by analytic integration for complex non-linear biochemical models.

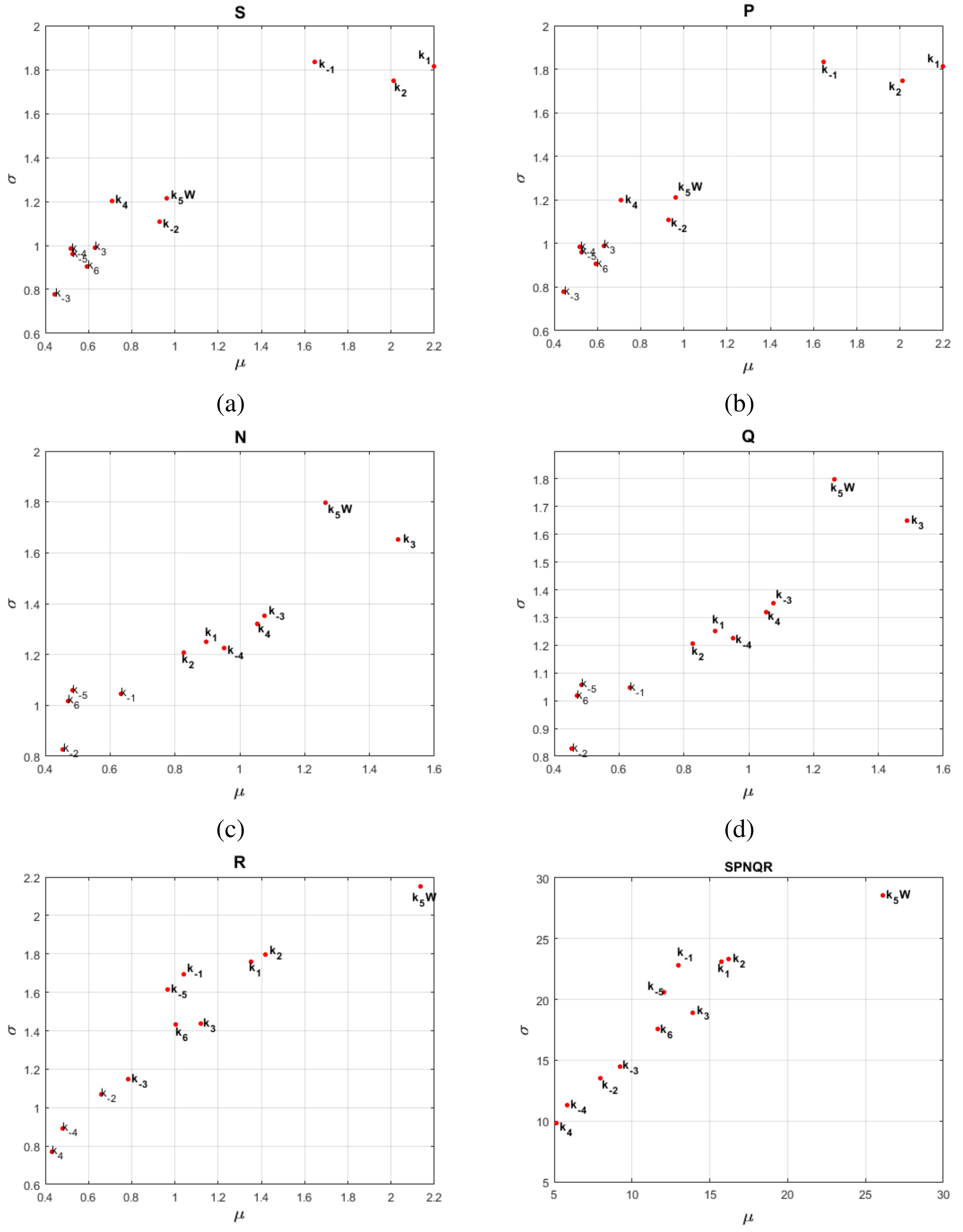
Example: GSA for enzyme reaction system

In this simulation study, the Morris sensitivity method and Sobol's sensitivity method are applied to an enzyme reaction system, which has been described in Section 4.1, in order to investigate parameter effects on model outputs. For the Morris screening sensitivity analysis, the nominal input condition (shown in Table 4.2) is used and 200 equally spaced sampling points along the whole reaction time (6000 seconds) are selected. The uncertainty ranges of model parameters are set to be 50% to 150% of their nominal values (given in Table 4.3) and parameters are assumed to follow uniform distribution. The number of levels p and the repetition number of the simulation γ are set to be 6 and 100, respectively. In order to avoid the cancellation of elemental effect, equation (C.1) is modified as:

$$EE_i(\boldsymbol{\theta}) = \sqrt{\left(\frac{f(\theta_1,\cdots,\theta_i+\Delta,\cdots,\theta_p)-f(\boldsymbol{\theta})}{\Delta}\right)^2} \tag{C.6}$$

The parameter effect on different model outputs by using the Morris screening method and the parameter ranking based on the mean value of elementary effects is shown in the following table and figure.

It can be seen that for state variable S and P parameters k_1, k_2 and k_{-1} are obviously the most important parameters. The effect of parameters on state variable N and Q are rather different, in which the most important parameters are k_3 and k_5W. For state variable R, k_5W is more important than other parameters. By considering all five model outputs, parameters k_1, k_2, k_3, and k_5W are the most important parameters.



(a)

(b)

(c)

(d)

(e)

(f)

Estimated mean and standard deviation of elementary effects of eleven model parameters for the five model outputs

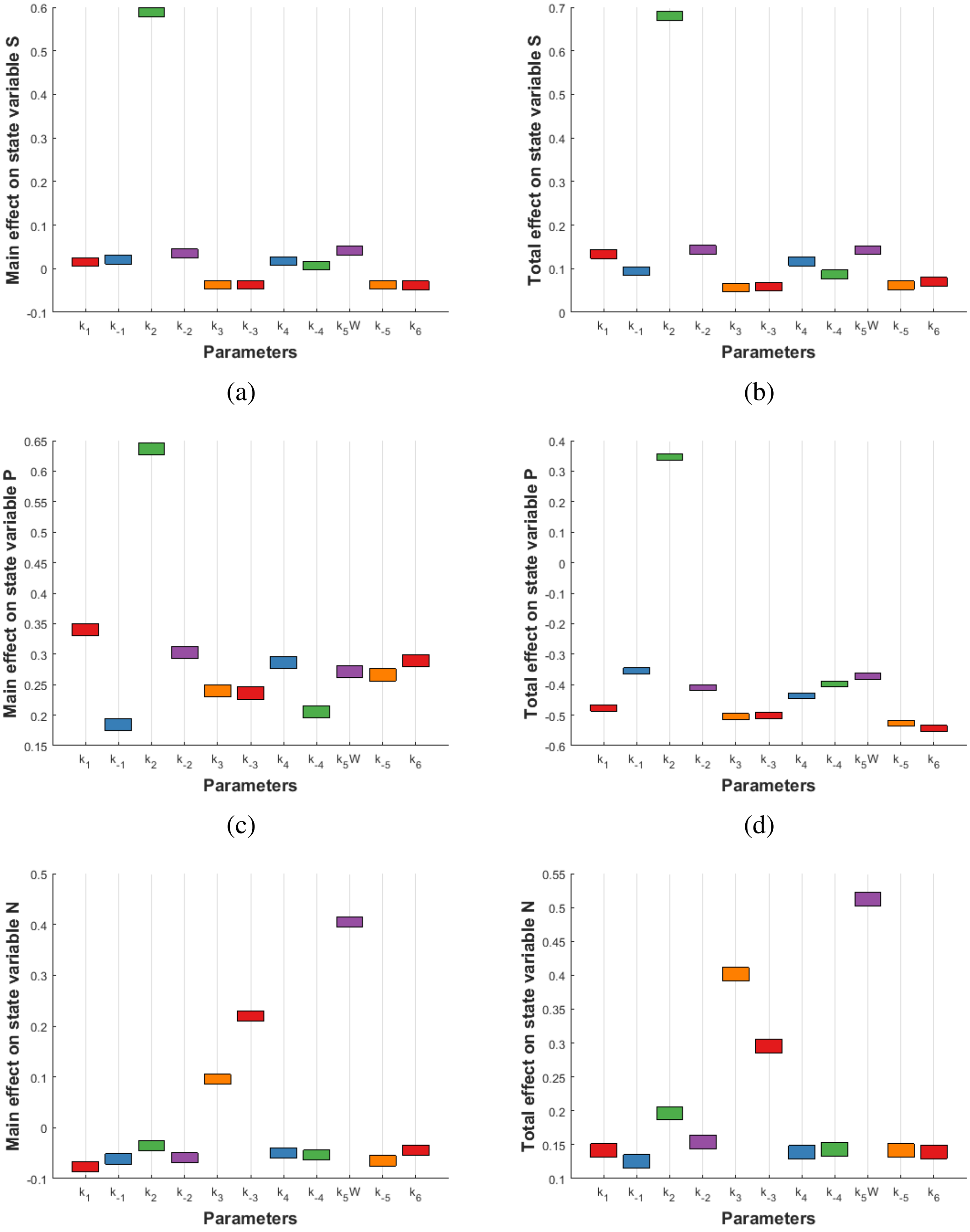
Parameter importance ranking based on the mean value of elementary effects

Model outputs	Parameter ranking (descent)
S	$k_1, \quad k_2, \quad k_{-1}, \quad k_5W, \quad k_{-2}, \quad k_4$
P	$k_1, \quad k_2, \quad k_{-1}, \quad k_5W, \quad k_{-2}, \quad k_4$
N	$k_3, \quad k_5W, \quad k_{-3}, \quad k_4, \quad k_{-4}, \quad k_1$
Q	$k_3, \quad k_5W, \quad k_{-3}, \quad k_4, \quad k_{-4}, \quad k_1$
R	$k_5W, \quad k_2, \quad k_1, \quad k_3, \quad k_{-1}, \quad k_6$
All five outputs	$k_5W, \quad k_2, \quad k_1, \quad k_3, \quad k_{-1}, \quad k_{-5}$

Then the Sobol's sensitivity algorithm is applied to investigate the parameter effect which is based on the analysis of variance. Simulation condition about parameters and input variables are the same as in Morris screening analysis. The Latin hypercube sampling strategy is employed where ten thousand samples for model parameter are selected for the analysis. The first order sensitivity indices and total sensitivity indices by considering different model outputs are shown in the following figure. Obviously, for state variable S and P, k_2 is more important than other parameters, while for state variable N, Q and R, the most important parameters are k_5W, k_{-3}, k_3 and k_2.

Robust sampling time design

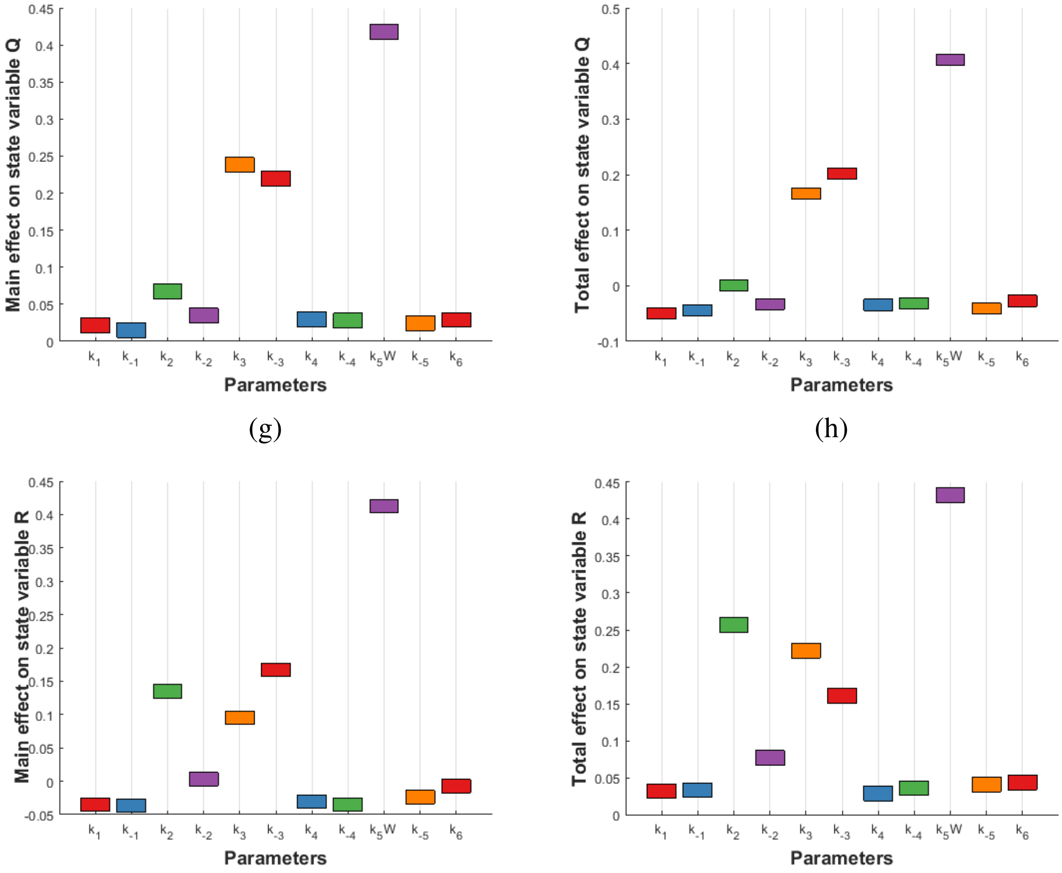
The first and most often used approach to overcome design problems related to parameter uncertainty is to design experiments in a sequential way by alternating parameter estimation and experimental design . Each estimation improves the knowledge of the system parameters and this knowledge can then be used to improve the quality of the next experiment to be performed. Many authors acknowledge the usefulness of this approach. However, some drawbacks are associated with this technique. Firstly, it might not be possible to perform multiple (sequential) experiments on the same system due to limitations in time or resources. Secondly, it is not guaranteed that the parameters will converge to the true values.


Main effect on state variable S
Total effect on state variable S
Main effect on state variable P
Total effect on state variable P
Main effect on state variable N
Total effect on state variable N
Parameters
k_1 k_{-1} k_2 k_{-2} k_3 k_{-3} k_4 k_{-4} k_5W k_{-5} k_6

(a) (b)

(c) (d)

(e) (f)



(g) (h)

(i) (j)

Parameter effect to model outputs based on the analysis of variance, including ‘main effect’ and ‘total effect’

A second approach to robust experimental design aims at determining the experiment that optimises the worst possible performance for any values of $\boldsymbol{\theta}$ belonging to the parameter domain $\boldsymbol{\Theta}$ (Flaherty et al., 2006; Körkel* et al., 2004; Rojas et al., 2007):

$$\boldsymbol{\phi}^* = \arg \max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\psi\left(FIM(\boldsymbol{\theta},\boldsymbol{\phi})\right)\right) \tag{C.7}$$

For a proposed design, find the parameters for which the scalar value of the FIM is the lowest, i.e. the worst possible obtainable information content for this specific design is determined. Next, find the design which maximizes the scalar value of the FIM with the worst parameter combination. For this technique, the prior information on the parameters is limited to the knowledge of the parameter domain, i.e. the upper and lower bounds of the parameters. No information on the probability distribution of parameters is necessary.

Instead of dealing with the parametric uncertainty with a worst-case maxi-min design strategy, an alternative way is to take account of parametric uncertainty by considering a prior distribution $p(\boldsymbol{\theta})$ of parameters in the design process, which leads to the Bayesian robust experimental design (BED) (Clyde, 2001; Murphy et al., 2003; Tulsyan et al., 2012; Vanlier et al., 2012). Bayesian experimental design is to quantify the statistical representation and treatment of the experimental design problem under Bayesian framework. Different from classical OED for nonlinear models which depends on the nominal parameter values, or maximin design that is based on the worst-case parameters, Bayesian designs are based around prior distribution of parameter estimates and their variance, rather than on chosen single-point values. Thus importantly, they can incorporate all previous scientific knowledge of model parameter estimates. A comprehensive review of Bayesian experimental design was given by Chaloner and Verdinelli (Chaloner and Verdinelli, 1995). It presents experimental designs in a unified decision-theoretic framework; therefore, different utility function can be specified, reflecting the purpose of a specific experiment, as a result different (alphabetic) design criteria become part of a single, coherent approach. In the Bayesian design framework, an experiment is said to be optimal if it satisfied (Huan and Marzouk, 2014; Kramer and Radde, 2010; Long et al., 2013; Ryan et al., 2014):

$$\boldsymbol{\phi}^*_{ED} = \underset{\boldsymbol{\phi}\in\boldsymbol{\Phi}}{\arg\max}\, \mathbb{E}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left(\psi\left(FIM(\boldsymbol{\theta},\boldsymbol{\phi})\right)\right) \tag{C.8}$$

where $\mathbb{E}$ represents the expected value of the scalar function ψ over all possible parameter values. The expectation is calculated with respect to the prior distribution of parameters $p(\boldsymbol{\theta})$, over the entire parameter space $\boldsymbol{\Theta}$ and the maximization is performed over the entire space $\boldsymbol{\Theta}$. Typically, numerical integration using a discretised version of the probability density function is used to evaluate the expectation. This solution is numerically quite burdensome if fine discretisation is used and many parameters are involved. Therefore, a multi-dimensional quadrature rule for approximating the multiple integral (over all parameters) can be applied. Another alternative would be to use Monte Carlo techniques to approximate the expectation, requiring however a large number of realizations before convergence can be achieved.

The connection between global sensitivity analysis and optimal experimental design has been discussed recently by Chu et. al. (Chu and Hahn, 2013). They investigated the consistency condition for applying optimal experimental design criteria to global sensitivity analysis results. They concluded that when a model is linear in model parameters, then the design based on the global sensitivity matrix is able to reduce to the conventional linear design based on the design matrix, which is:

$$\lim_{\Delta\boldsymbol{\theta}\to 0} \psi\left(\mathbf{S}_G^T\mathbf{S}_G\right) = \psi\left(\mathbf{S}_G^L\mathbf{S}_L\right) \tag{C.9}$$

One should note that not all global sensitivity analysis techniques satisfy the consistency condition and the author selects four different sensitivity analysis functions for the optimal experimental design. In this work, two different global sensitivity analysis techniques are integrated to the optimal experimental design. The first one is given by the mean of the local sensitivity over parameter uncertainty region, which can be expressed as:

$$s_i = \mathbb{E}\left[\frac{\partial X}{\partial \theta_i}\right] \tag{C.10}$$

The expectation of local sensitivity over all possible parameter values represents the average effect of a parameter. The other sensitivity analysis technique is based on the mean of squared sensitivity values which can avoid the cancellation of the negative effect of parameter sensitivities, the formulation of which is:

$$s_i = \sqrt{\left\{\frac{\mathbb{E}\left[\left(\frac{\partial X}{\partial \theta_i}\right)^2\right]}{var\left[\frac{\partial X}{\partial \theta_i}\right]} - 1\right\}} \tag{C.11}$$

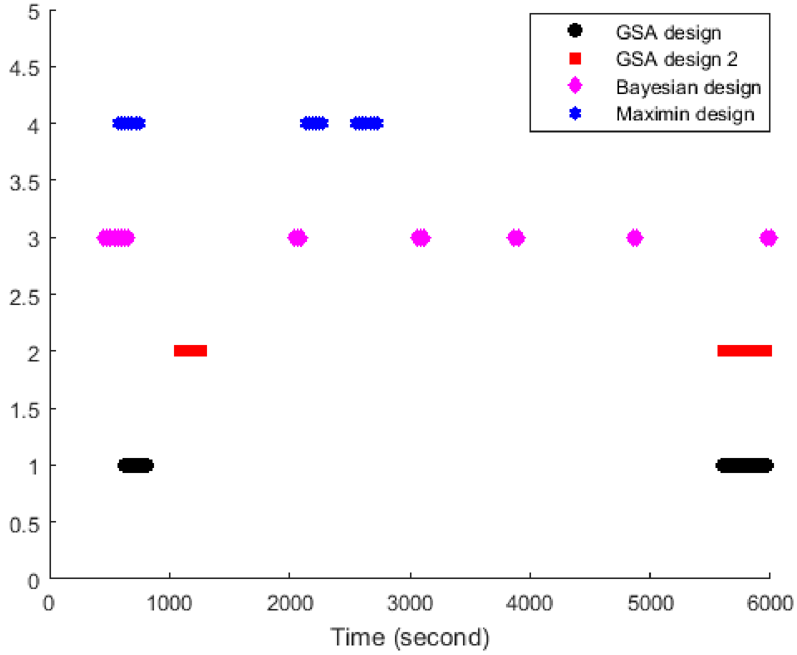
In this simulation, three different robust design methods are applied and compared for the sampling time design of the enzyme reaction system described in Section 4.1. Parameters k_2, k_{-3} and k_5W are selected in the design framework. In all three different experimental design simulations, the parameter uncertainty range is set to be 0.5-1.5 times of nominal parameter values in Table 4.3 and nominal condition of input values in Table 4.2 is applied. For the maximin robust design, the Powell's method is applied for the optimal sampling time selection and the particle swarm algorithm is implemented to find the worst parameter values that will lead to the least informative data. In the Bayesian sampling time design, the Monte Carlo sampling strategy is employed to select ten thousand random parameter sets. The same method is used in the global sensitivity analysis for the selection of possible parameter values. In all three experimental design strategies, the D-optimal design criterion is applied and the objective is to find 20 best sampling time points that can best facilitate parameter estimation when parameter uncertainties are considered. The simulation result is shown as follows:

Robust sampling time design with different RED techniques

RED methods	Sampling time points (sec)	D-values
Maximin design method	330:30:510 3300:30:3450 4800:30:4980	8.84e-8
Bayesian design method	450:30:660 2040:30:2100 3060:30:3120 3870 3900 4860 4890 5970 6000	6.02e-6
GSA based design (C.10)	630:30:810 5610:30:5970	1.47e-5
GSA based design2 (C.11)	1080:30:1260 5610:30:5970	0.0012

For the maximin robust design, the final obtained parameter values for the three selected parameters are $k_2 = 150$, $k_{-3} = 300$ and $k_5W = 7500$. By using this design method, the data information is maximized even if the worst parameter sets are chosen. The exact sampling time points are selected at the start and middle stage of the reaction while data information in the late stage of reaction is ignored. For the Bayesian robust sampling time design, the sampling time is loosely distributed along the whole reaction time and the averaged data information is maximized. The relatively equally spaced sampling strategy is suggested. In other words, data information at each available sampling point has equal importance when model parameters contain large uncertainties. However, this sampling time strategy might not be informative for a specific parameter set. The GSA based experimental design can lead to more informative data. The exact sampling time points can be divided into two groups, in which sampling time points should be selected at the start of the reaction and at the end of the reaction.

In order to compare various design results, we select 27 different parameter sets for k_2, k_{-3}, and k_5W, where each parameter can be chosen as 0.5, 1 and 1.5 times of their nominal values. The D-optimal values of different parameter sets with their corresponding optimal sampling strategies are compared as follows:



Sampling time distribution obtained from different RED methods

Comparison of D-values among different RED methods with various parameter sets, $\ln(\det(FIM))$ is used as the metric

RED methods	best D-value	mean of D-values	standard deviations
Maximin design method	-34.16	-38.44	2.87
Bayesian design method	-33.80	-38.22	2.61
GSA based design (C.10)	-50	-52.49	2.07
GSA based design2 (C.11)	-33.75	-41.75	4.17

It can be seen that the results from GSA-based designs are worse than those from maximin and Bayesian designs. When parameter sensitivities are averaged during GSA, it is likely to find only a local optimal design. The performance of GSA based design is only good when parameter estimates are very close to true values and the sampling strategy might result in non-informative data with other parameter sets. The standard deviation value also shows that this kind of design is sensitive to parameter values. The maximin design and Bayesian design, however, can provide higher D-values and small deviations, which indicate that these robust design methods are less sensitive to the change of parameter values and can provide in average higher data information than local design or GSA based design. These two methods are useful when model parameters are very uncertain. In addition, the maximin design is computationally less expensive than that of the Bayesian design in this particular case. Using the maximin design method, the worst parameter set can be determined through less than 100 iterations while the Bayesian design requires OED optimisation for a large number of different parameter sets (1000 in this case).