Comprehensive experimental design for chemical engineering processes: a two-layer iterative design approach

Hui Yu\textsuperscript{a}, Hong Yue\textsuperscript{a}, Peter Halling\textsuperscript{b}

\textsuperscript{a}Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK
\textsuperscript{b}WestCHEM Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow G1 1XL, UK

Abstract

A systematic framework for optimal experimental design (OED) of multiple experimental factors is proposed to support data collection in chemical engineering systems with the purpose to obtain the most informative data for modeling. The structural identifiability is firstly investigated through a combined procedure with the generating series method and the identifiability tableau. Next the parameter estimability is analyzed via the orthogonalized sensitivity analysis in order to identify crucial and identifiable model parameters. Traditionally OED treats separate problems such as the choice of input conditions, the selection of variables to measure, and the design of sampling time profile. A new OED strategy is proposed that optimizes these interdependent factors in one 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 optimization of input intensities is obtained through stochastic global searching. In this way, the multi-factor optimization problem is solved through the integration of a stochastic method, for the upper layer, and a deterministic method, for the lower layer. Case studies are conducted on two biochemical systems with different complexities, one is an enzyme kinetically controlled synthesis system and the other one is a lab-scale enzymatic biodiesel production system. Numerical results demonstrate the effectiveness of this double-layer OED optimization strategy in reducing parameter estimation uncertainties compared with conventional approaches. 

Keywords: optimal experimental design (OED), multi-factor optimization, input conditions, sampling time profile, measurement set selection, chemical reaction systems.

1. Introduction

Mathematical models are widely used in chemical and biochemical process engineering since the mathematical representation enables to reproduce real dynamic processes in a simulation environment (Baltes et al., 1994; van Riel, 2006; Bogacka et al., 2011; Villaverde et al., 2014). 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 control and predict the future behavior of systems and to test hypotheses (Phair, 1997; Peleg et al., 2002; Fages et al., 2004; de Brauwere et al., 2009; Liepe et al., 2013; Yu et al., 2015). A typical modeling procedure consists of several important steps (Franceschini and Macchietto, 2008), as shown in
Fig.1. Once one or several candidate models are proposed from prior knowledge, it is necessary to investigate if it’s possible to obtain unique solutions for model parameters under ideal conditions of noise free observations and error-free model structures, if not, alternative models need to be proposed. For those structurally identifiable models, parameter sensitivity analysis and estimability analysis are required which will help to make model calibration more specific on those key parameters, whereas non-important parameters can be kept on their nominal values or even be removed so as to reduce the model complexity. The most suitable model can then be determined through fitting with experimental data, which is referred to as model calibration in Fig.1. The established model needs to be further validated using experimental data.

Model development of process systems is normally an iterative process that includes steps on data collection, model selection, model calibration and model validation until a satisfactory model is obtained with acceptable predictive capabilities. It requires large amounts of experimental data at all modeling steps. For chemical reaction systems, a typical method is to represent reactions into a set of coupled differential equations based on certain conceptual framework, e.g. mass-action laws. The reactants and products involved in the reaction network are therefore interconnected with kinetic parameters, whose values are generally unknown \textit{a priori}. One of the main goals in model building is then to estimate those unknown parameters based on experimental data. However, measurement of process variables especially reactants is restricted by many factors such as sensor technology, operation constraints, limited time and budget, etc. Constraints on inputs can also
affect implementation of experiments. What’s more measurement data are inevitably contaminated with experimental noise. The lack of sufficient and accurate measurement data makes model development a challenging task especially when the system is high dimensional, nonlinear with poorly understood dynamics like many complex biological or biochemical networks.

In data-based model development, it is essential to obtain high quality and informative measurement data with less experimental efforts if possible. Therefore, modern experimental design techniques play important roles in model building process at various stages. The purpose of optimal experimental design (OED) is to devise necessary experiments that are most
likely to generate data that will best facilitate the identification of model
structure and the determination of model parameters (Faller et al., 2003).
Typically an OED problem can be formulated as a dynamic optimization
problem with respect to the design factors of interest. The major objective
of OED is to maximize the data information through a measure of certain
scalar function of Fisher information matrix (FIM) (Balsa-Canto et al., 2008).
The design factors can normally be classified into two groups, i.e., the input
design factors and the observation design factors. The former determines
the stimulation and control actions, e.g., the initial conditions, the external
time-dependent input conditions. These input factors will change the system
dynamics. The latter is to determine which to measure, when to measure and
where to measure, for example, design of sampling time profiles and design
of measurement set selection. Here the measurement set refers to the choice
of variables to be measured.

Various OED methods have been developed for data-based modeling of
chemical, biological and wider systems aiming at individual experimental fac-
tors such as a factor in input settings (Chianeh et al., 2011; Yue et al., 2013)
or a factor in observation design (Kutalik et al., 2004; Brown et al., 2008;
Asyali, 2010; He et al., 2010). When a single factor is determined individually
through OED, the design result and the overall information contained in the
experimental data are dependent on other factors that are not included in
the design. If those non-designed factors are not properly chosen, the quality
of the experimental data cannot be guaranteed. For a dynamic system to be
modeled, such dependence among experimental factors exist between the in-
put factors, the measurement factors, and the interaction between them. To
reduce the uncertainty in single factor design and increase the data quality, a more effective OED should support optimization of multiple experimental factors in a systematic way. Very few works have been reported on how to tackle OED of multiple experimental design factors mainly because it is very difficult to obtain OED solutions for multiple design variables for a complex nonlinear system, not to mention the system constraints and the operational constraints that need to be considered in optimization. One option for multi-factor OED is to implement the optimization of multiple experimental parameters through a sequential design strategy in which each single factor is designed iteratively and the interested experimental factors are updated at each iteration, however, this method is computationally rather cumbersome and does not necessarily assure the global best design.

An OED problem including multiple experimental factors normally contains a large number of design variables and has multiple local maxima/minima (Banga et al., 2002), for which the commonly used gradient-based optimization methods may only converge to local optima. Various global optimization techniques have been developed to solve complex OED problems with the purpose to obtain global optima and improve the convergent speed, see (Banga et al., 2005; Catania and Paladino, 2009; Ruffio et al., 2012) for example. Most global optimization techniques are population-based requiring a large number of calculations of model equations and objective functions. The computational load is increased exponentially with the increase in the number of design variables. This makes OED of multiple experimental factors computationally demanding. In this work, we aim to develop a framework to conduct OED of multiple experimental factors in an integrated, compu-
tationally efficient environment so that the data collected from the designed
experiments contain rich information for modeling. This framework will sup-
port modeling related tasks such as simulation of complex dynamic systems,
fundamental system analysis that are crucial for OED and parameter estima-
tion, e.g. parametric sensitivity analysis, structural identifiability analysis,
parameter estimability analysis, OED of multiple input factors and observa-
tion factors, assessment of OED results, etc.

The remaining of the paper is organized as follows. Section 2 presents
preliminaries on least-square parameter estimation and OED relevant analy-
sis such as parametric sensitivity analysis, structural identifiability analysis
and parameter estimability analysis. In Section 3, development of several key
OED problems are presented on single experimental factors including input
intensities, measurement set selection and sampling time profile, individu-
ally, using different optimization strategies. A novel integrated observation
design is proposed in Section 4, where the measurement set selection and
the sampling profile are determined simultaneously. In Section 5, an itera-
tive two-layer design is proposed for integrated design of input factors and
observation factors. OED on two case study systems, an enzyme reaction
system and a lab-scale enzymatic biodiesel production system, are simulated
and discussed in Section 6. Finally, conclusions and discussions are made in
Section 7. Details of case study models are given in Appendix.

2. Preliminaries on relevant methods

Consider a general nonlinear dynamic model with \( n \) state variables, \( p \)
parameters and \( m \) output variables, the state and output can be described
by a set of ordinary differential equations (ODEs) and algebraic equations:

\[ \dot{X}(t) = f(X(t), \theta), \quad X(t_0) = X_0, \]  
\[ Y(t) = h(X(t), \theta) + \xi(t). \]  

where \( f(\cdot) \) is a set of state transition functions of the system dynamics which are assumed to be continuous and first-order derivative; \( X = [x_1, x_2, \cdots, x_n]^T \) \( \in \mathbb{R}^n \) denotes the vector of \( n \) state variables with initial condition \( X_0 \); \( \theta = [\theta_1, \theta_2, \cdots, \theta_p]^T \in \mathbb{R}^p \) is the vector of \( p \) model parameters; \( Y \in \mathbb{R}^m \) is the measurement output vector with \( m (m \leq n) \) measurement variables; \( h(\cdot) \) is the measurement output function, normally used for selecting which variables to be measured. \( \xi \) 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 contaminate the observations always exist. Most often the measurement error is assumed to be a zero mean, Gaussian noise.

2.1. Least-square parameter estimation

Model parameters can be estimated using collected measurement data. When the system model is linear in parameters or can be transformed to be linear in parameters, a widely used method for parameter estimation is the (weighted) least-square estimation, where the problem is formulated as

\[ \hat{\theta} = \arg \min_{\theta \in \Theta} J(\theta) \]

\[ = \arg \min_{\theta \in \Theta} \sum_{l=1}^{N} \left( Y(t_l) - \hat{Y} \left( \hat{\theta}, t_l \right) \right)^T \cdot Q^{-1} \cdot \left( Y(t_l) - \hat{Y} \left( \hat{\theta}, t_l \right) \right), \]  

where \( Y \) and \( \hat{Y} \) 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. Assuming all observation variables can be measured independently and characterized by the variance of $\sigma_j^2$, the measurement error covariance matrix is written as $Q = \text{diag}[\sigma_1^2, \ldots, \sigma_m^2]$.

The adequacy of the model and the parameter significance can be assessed by evaluating the output residuals through statistical tests. The method based on joint confidence regions between parameters is widely used to evaluate the estimation quality (Franceschini and Macchietto, 2008). The confidence region can be determined based on the following cost function:

$$\left\{ \mathbf{\theta} : J(\mathbf{\theta}) \leq \left( 1 + \frac{p}{N-p} F_{p,N-p}^{1-\alpha} \right) \times J(\hat{\mathbf{\theta}}) \right\},$$

(4)

where $F_{p,N-p}^{1-\alpha}$ is the upper $\alpha$-critical level of $F$ distribution with $p$ and $(N-p)$ degrees of freedom; $\alpha$ is a positive real number between 0 and 1. However, for a nonlinear model, $J(\mathbf{\theta})$ is not a quadratic function with respect to $\mathbf{\theta}$, a linearization approximation is made by Taylor expansion around the estimated parameters $\hat{\mathbf{\theta}}$. The confidence region can then be approximated as (Ljung, 1987)

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

(5)

where

$$\mathbf{V} = 2 \times \frac{J(\hat{\mathbf{\theta}})}{N-p} \times \mathbf{H}(\hat{\mathbf{\theta}})^{-1}, \quad \mathbf{H}(\hat{\mathbf{\theta}}) = \frac{\partial^2 J}{\partial \mathbf{\theta} \partial \mathbf{\theta}^T}$$

(6)

Here $\mathbf{V}$ is the parameter estimation error covariance matrix which is used as the cornerstone to measure parameter estimation uncertainty. $J(\hat{\mathbf{\theta}})/(N-p)$ is an approximation of residual variance. $\mathbf{H}$ is the Hessian matrix. The confidence interval of a single parameter $\theta_i$ can be determined by

$$\delta_i = \pm t_{N-p}^{\alpha} \times \sqrt{V_{ii}}$$

(7)

where $t_{N-p}^{\alpha}$ is the student distribution with $(1 - \alpha)$ confidence level and $(N - p)$ degrees of freedom. In later discussions, the formulation in (4)-(7)
will be used to produce confidence intervals to assess uncertainty in parameter estimation.

2.2. Structural identifiability analysis

As a key step and normally the initial step in parameter estimation scheme, structural identifiability analysis is performed to figure out whether it is possible to obtain unique parameter values for the candidate model structure from the data. If the parameters can be uniquely estimated from noise-free experimental data, then the model is said to be structurally identifiable. Consider the general dynamic model in (1) - (2), if

\[ \forall \theta_1, \theta_2 \in \mathbb{R}^p, \ h(X(t), \theta_1) = h(X(t), \theta_2) \iff \theta_1 = \theta_2, \tag{8} \]

the parameters in \( \theta \) are said to be globally identifiable. If the condition holds only for a neighbourhood of \( \theta^* \) in the parameter space which is given by

\[ \forall \theta_1, \theta_2 \in \{ \theta \in \mathbb{R}^p \mid \| \theta - \theta^* \| < \delta \}, \]

\[ h(X(t), \theta_1) = h(X(t), \theta_2) \iff \theta_1 = \theta_2, \tag{9} \]

the parameters \( \theta \) are said to be locally identifiable (McLean and McAuley, 2012). A number of methods have been developed to check the structural identifiability of nonlinear models such as Taylor series expansion approach (Pohjanpalo, 1978), generating series method (Walter and Lecourtier, 1982), local state isomorphism (Vajda et al., 1989) differential algebra algorithm (Ljung and Glad, 1994), or check the structural identifiability of the linearized part of the nonlinear model (Ben-Zvi et al., 2006). With the development of symbolic computational tools, the power series expansion methods that include the Taylor series expansion approach and the generating series method have been developed for structural identifiability analysis.
The basic idea of the Taylor series expansion approach is that the observations of the system under consideration have unique analytic representations with respect to time, and therefore their derivatives with time are also represented uniquely. Thus it is possible to represent the observations by using Maclaurin series expansion, written as

\[ y_i(\theta, t_0 + \Delta t) = y_i(\theta, t_0) + \frac{dy_i}{dt} \Delta t + \frac{1}{2} \frac{d^2y_i}{dt^2} (\Delta t)^2 + \ldots \]  

(10)

where \( \Delta t \) is a small time increment. The uniqueness of those Taylor series coefficients in (10) can guarantee the structural identifiability of the model.

With the generating series approach, the observations are expanded with respect to time and inputs. This method is refined to state models which are linear in the inputs, given as follows:

\[ \dot{X}(t) = f(X(t), \theta) + \sum_{i=1}^{n_u} g_i(X(t), \theta) u_i(t) \]  

(11)

where \( u_i \) stands for input factors, \( n_u \) is the number of input factors, and \( g_i \) is the corresponding coefficient for \( u_i \). The observations in (2) can be expanded in such a way that the series coefficients are \( h(X_0, \theta) \) and its Lie derivatives, \( L_{f_0} h, L_{f_1} h, \ldots, L_{f_{nk}} h \), where \( L_{\phi} h (X_0, \theta) = \sum_{j=1}^{n_u} g_j (X_0, \theta) \cdot \frac{\partial}{\partial x_j} h (X_0, \theta) \).

Similar to the Taylor series approach, the structural identifiability problem is transformed into the determination of power series coefficients, the unique value of which provides a sufficient condition of structurally identifiable model. However, it should be noted that there is no upper bound for the number of derivatives that needs to be calculated for nonlinear models. For nonlinear systems with a large number of parameters the calculation of power series coefficient is a computational cumbersome work.

In this work, the identifiability tableau method proposed in (Balsa-Canto et al., 2010) is used for structural identifiability analysis. The identifiability
tableau is constructed to represent the non-zero elements of the Jacobian matrix of those power series coefficients on model parameters. Some model parameters can be obtained directly from solving simple algebraic equations. With the obtained model parameter values, the identifiability tableau can be reduced and eventually minimized. The analysis of the remaining parameters will be conducted in a sequential procedure. The structural identifiability of a model parameter depends on the existence of the solution of that parameter. More details on identifiability tableau can be found in (Balsa-Canto et al., 2010; Chis et al., 2011).

2.3. Parameter estimability analysis

For a structurally identifiable model, its unknown parameters may still not be estimable in practice (also called practical identifiability) due to several reasons: (i) the experimental data for parameter estimation are sparse and noisy, or contains inadequate information due to poorly designed experiments; (ii) some unknown parameters have very little influence on model outputs, i.e., of low parametric sensitivities; (iii) the effect of some parameters to the model prediction can be compensated by other parameters, i.e., high correlations exist between parameters to be estimated.

Practical identifiability analysis is in general a discrete (combinatorial) non-convex optimization problem. Exhaustive search and genetic algorithms are the most widely used methods to get the solution. However, for nonlinear dynamic systems with a large number of parameters, these methods are computationally too expensive. Methods of approximations and relaxations of the original optimization problem have been developed and applied to evaluate practical identifiability. These include but not limited to the collinearity
index (Brun et al., 2001), the relative gain array (Sandink et al., 2001),
the Hanken singular value (Sun and Hahn, 2006), orthogonalization based
methods (Yao et al., 2003), optimization methods that rely on the Fisher
information matrix, and methods with repeated parameter estimation. In
this work, an orthogonalization based method (Yao et al., 2003) will be used
for practical identifiability analysis. This method is based on the measure
of orthogonal parameter sensitivities. In another word, the parameter pair
correlations have been removed from the original local sensitivity matrix and
the measurement is focused on the independent parameters.

2.4. Parametric sensitivity analysis

Parameter sensitivity analysis is a method used to examine how sensitive
the system output is in response to variations in model parameters. The
parametric local sensitivities can be described by

\[ \dot{S} = \frac{\partial f}{\partial X} \cdot \frac{\partial X}{\partial \theta} + \frac{\partial f}{\partial \theta} \]  

where \( S = \frac{\partial X}{\partial \theta} = [s_{ij}] \in \mathbb{R}^{n \times p} \) is the parameter local sensitivity matrix,
\( s_{ij} = \frac{\partial x_i}{\partial \theta_j} \); \( \frac{\partial f}{\partial X} \in \mathbb{R}^{n \times n} \) is the Jacobian matrix, and \( \frac{\partial f}{\partial \theta} \in \mathbb{R}^{n \times p} \) is the paramet-
ric Jacobian matrix. The state differential equations in (1) and the sensitivity
differential equations in (12) can be solved simultaneously through the direct
differential method (Atherton et al., 1975). To remove the effects of model
parameters that are likely to have values at different scales, normalized sen-
sitivities, \( \bar{s}_{ij} = \frac{\partial x_i}{\partial \theta_j} \cdot \frac{\theta_j}{x_i} \), are sometimes used for comparison of parameter
sensitivities. The corresponding normalized sensitivity matrix is \( \bar{S} = [\bar{s}_{ij}]_{n \times p} \).
The overall effect of parameter \( \theta_j \) to all state variables can be calculated by
a norm of local sensitivities such as
\[ OS_j = \frac{1}{N} \sqrt{\sum_{i=1}^{n} \sum_{l=1}^{N} s_{ij}^2(t_l)} \]  

(13)

The sensitivity analysis results can be used to find key parameters that have significant impacts to system behavior, to assist model simplification or used in gradient based optimization process for parameter estimation. In model based OED, the parametric sensitivity matrix is taken to construct the FIM. Therefore, parameter sensitivity plays an indispensable role in parameter estimation, parameter identifiability analysis and experimental design.

3. OED for single experimental factors

3.1. Fisher information matrix and design criteria

The task of model-based OED for parameter estimation is to determine the values of experimental variables so that the predicted measurement data information is optimized. Denoting the design factors which characterize the experiment into a vector \( \zeta \), the FIM can be locally written as

\[ \text{FIM} (\theta, \zeta) = S(\theta, \zeta)^T \cdot W \cdot S(\theta, \zeta) \]  

(14)

where the weighting matrix \( W \) is normally taken to be \( Q^{-1} \) for the 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, \( Q^{-1} \), in the FIM indicates that data with 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, $\Sigma$, under the assumption of unbiased parameter estimation and uncorrelated additive white measurement noise (Ljung, 1987).

The OED problem can be cast as minimization of a proper measure of the parameter error covariance matrix, which can be approximated as the inverse of FIM, i.e.

$$\zeta^* = \arg \min_{\zeta \in \Omega} \Phi \left( \left( \text{FIM} (\theta, \zeta) \right)^{-1} \right),$$

(15)

where $\Omega$ is the admissible space of the design factors, $\Phi (\cdot)$ represents a function to scalarize the inverse of FIM. The most commonly used design criteria in OED are A-optimal, D-optimal, E-optimal, and modified E-optimal designs, in which the scalar measures are closely related to the shape, size and orientation of parameter estimation confidence intervals. The design focus of these scalar design criteria are different from each other due to the different features taken from the FIM. No single design criterion can be applicable to all design problems or suitable for all systems. For a given dynamic system, one particular optimization criterion may be superior to others; but this does not necessarily mean that this criterion plays well in other designs. Therefore, it is recommended that different criteria should be tried and compared in a standard experimental design.

For most chemical and biochemical reaction systems, the OED for parameter estimation can be put into two categories, i.e. input design on manipulation of input variables, and observation design such as design of sampling time profile and selection of measurement variables. For a given
dynamic system, the change in input will change the dynamic response. This means during the OED process, for each value taken for an input factor, the full dynamic response profile needs to be calculated. On the other hand, in the design of observation variables, the dynamic response is determined by the specific input condition, only one calculation of the dynamic response is required for the optimization process. For this reason, the experimental design formulation and the optimization processes for the design of input factors and for the design of observation factors can be quite different.

3.2. Input intensity 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 nonlinear programming dynamic optimization problem by approximating the time-varying inputs with discrete form of inputs. The problem can then be solved by direct dynamic optimization methods such as the sequential methods and the simultaneous methods (Biegler et al., 2002).

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 OED problem is formulated as the general form in (15), in which the design factors are the initial input intensities, i.e., \( \mathbf{\zeta} = \mathbf{X}_0 \). It should be noted that only those elements in \( \mathbf{X}_0 \) that need to be designed are included in the OED, other initial conditions are kept at the values according to the system mechanism and operating conditions.
Since the response of a dynamic system will change following the change in inputs, the FIM is also changed and needs to be calculated for each choice of the input. Numerically this will involve integration of ODEs in (1) being implemented many times during the optimization process.

This input design is in general a non-convex optimization 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 optimization (PSO) algorithm is chosen, which has not been used in previous multi-input OED.

3.3. Measurement set selection

Collecting measurement data with rich information for modeling could be cost expensive and time-consuming, especially for complex biological or biochemical systems. The aim of OED on measurement set selection is to find a necessary or a minimum set of variables to be measured such that the selected measurement variables are most useful or discriminating for parameter estimation. From the system development point of view, another benefit from optimized measurement set selection is that the design results may indicate missing measurement of variables that are actually crucial to modeling. Necessary measurement can then be added to the sensing system. In this work, it is assumed that each state variable can be measured independently. For some circumstances where only combination of states can be measured, similar design can still be applied since the importance of the combined measurement of interest can be easily determined from the ranking (and the weighting) of each individual state after the OED.

Assuming that the measurement set is selected from the full set of the
state variables, the measurement set selection problem can be formulated as follows (Flaherty et al., 2006):

$$\xi \equiv \begin{cases} x_1 & \cdots & x_n \\ \lambda_1 & \cdots & \lambda_n \end{cases}$$

$$\xi^* = \arg \min_{\lambda \in \Omega} \Phi \left( \left( \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \lambda_i S_i^T S_i \right)^{-1} \right)$$

$$s.t. \quad \lambda_i \in \{0, 1\}, \quad 1^T \lambda = n_{sel}$$

where $1$ is a column vector comprised of ones in all its entries; $\lambda = [\lambda_1, \cdots, \lambda_n]^T$, in which $\lambda_i$ is the non-negative weight factor for $x_i$ that can be chosen as either 1 or 0; $n_{sel}$ is the total number of measurement variables to be used. After the OED, those state variables with weighting factor values to be 1 are selected to form the measurement set.

### 3.4. Sampling time profile design

The target of optimal design of sampling time profile(s) is to determine the sampling time points that will enable most informative data collection at those points. The design problem can be set up as to choose certain number of sampling points along the measurement states, which, in principle, is an infinite dimensional non-convex dynamic optimization problem hard to solve. To tackle this difficulty, the sampling time profile design can instead be formulated as a discrete optimization problem. The available measurement variables are defined \textit{a priori}, also the total number of sampling points is given for each measurement variable, and the OED is performed to find the best combination of a subset of the data points from the whole set.
Similar to the OED of measurement set selection, the optimal design problem of sampling time profile can be formulated as follows:

\[
\xi = \begin{cases} 
    t_1 & \cdots & t_N \\
    \omega_1 & \cdots & \omega_N 
\end{cases}
\]

\[
\xi^* = \arg \min_{\omega \in \Omega} \Phi \left( \left( \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \omega_i S(t_i) S(t_i)^T \right)^{-1} \right) 
\]

\[
\text{s.t. } \omega_i \in \{0, 1\}, \ 1^T \omega = N_{sp}
\]

where \( \omega = [\omega_1, \cdots, \omega_N]^T \) is the weighting vector for all the available measurement points in time horizon. \( N_{sp} (\leq N) \) is the total number of sampling points to be selected. Here it is assumed that the same sampling time profile is applied to all considered measurement variables. One should note that time resolution should be small enough so that the optimal sampling time solution are included in the predefined sampling time set.

### 4. Integrated observation design

We start from OED of observations by fixing the input experimental factors. The observation design of measurement strategies include but are not limited to the measurement set selection and the sampling time profile design. Compared with the input experimental design, one big advantage in design of measurement factors is that when the input (stimulation/perturbation) is fixed, the dynamic response of the system is also determined, in other words, the candidate pool of the available measurement information is provided. The observation design is mainly to find a strategy that can pick up the most informative data from the available measurement data.
As can be seen from Sections 3.3 and 3.4, the design of sampling time profile and the design on measurement set to be selected are handled separately. 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 could be correlated 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 is to go through an iterative procedure to design the two experimental factors, in each iteration only one factor is optimized 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.

Here we propose to combine the measurement set selection and the sampling time profile design into one single optimization problem. This idea is inspired by the fact that the two optimization problems share a similar formulation as in (16) and (17), and only one integration of the state variables is required during the optimization design under given input. The integrated observation design is represented as the following optimization problem.

\[
\begin{align*}
\xi &= \left\{ \begin{array}{ccc}
t_1 & \cdots & t_{N\times n} \\
\omega_1 & \cdots & \omega_{N\times n}
\end{array} \right\} \\
\xi^* &= \arg\min_{\omega \in \Omega} \Phi \left( \left( \sum_{i=1}^{N\times n} \frac{1}{\sigma_i^2} \omega_i S(t_i)S(t_i)^T \right)^{-1} \right) \\
\text{s.t.} & \quad \omega_i \in \{0, 1\}, \quad 1^T\omega = N_{ssp}
\end{align*}
\]

Here the number of the integrated weighting factors is extended to \(n \times N\)
for the system with $n$ state variables and the data length of $N$, i.e., $\omega = [\omega_1, \omega_2, \cdots, \omega_{n \times N}]^T$. Each $\omega_i$ stands for the importance of one measurable state variable at a particular time point. $N_{ssp}(\leq n \times N)$ is the total number of sampling points to be selected for all the state variables at all chosen time points. The design problem as formulated in (18) is an integer programming problem which can be solved by exhaustive search if the number of $n \times N$ is relatively small. For a design that contains a large number of weighting factors, the optimization problem in (18) can be further relaxed to an approximate continuous optimization problem (Yue et al., 2008; He et al., 2010), which is given as follows.

$$\xi^* = \arg\min_{\omega \in \Omega} \Phi \left( \left( \sum_{i=1}^{N \times n} \sigma_i^2 \omega_i S(t_i)^T S(t_i) \right)^{-1} \right)$$

(19)

$$s.t. \sum_{i=1}^{N \times n} \omega_i = 1, \ \omega_i \geq 0$$

The weighting term $\omega_i$ is relaxed to a continuous variable taking values between $[0, 1]$. In this way, the the optimal solution provides a lower bound for the original integer optimization problem. At each sampling time point, the FIM for involved state variables is a positive definite matrix. Therefore, the continuous optimization problem in (19) can be converted into a convex optimization problem by employing different scalar design criteria. For instance, taking the D-optimal design criterion, the observation design problem can be easily transformed into a convex optimization problem that can be solved by local optimization methods such as the Powell’s quadratically
convergent method (Kutalik et al., 2004) or the interior-point method. When the A-optimal or E-optimal design criterion is applied, problem (19) can be transferred into an equivalent semi-definite programming (SDP) problem. The E-optimal observation design formulation is written as follows.

\[
\begin{align*}
\min & \quad -t \\
\text{s.t.} & \quad \sum_{i=1}^{n \times N} \frac{1}{\sigma_i^2} \omega_i S_i^T S_i \succ t I \\
& \quad \omega_i \succ 0, \forall i; \quad 1^T \omega = 1
\end{align*}
\]  

The optimization problem in (20) can be conveniently solved by available computational tools such as the 'SeDuMi' software. When the gradient-based optimization method is used to solve the problem, the derivative of the objective function over the weights is much easier to calculate than the direct derivative over time and state variables. With this integrated design, the sampling time profile and the measurement set are simultaneously determined through a single-objective optimization.

5. Iterative double-layer design of both observation and input

In a systematic experimental design, those major experimental conditions such as the input perturbations and the measurement strategy should be considered in an integrated design framework. This integrated optimization problem can be handled through a sequential 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 optimization problem as discussed in Section 3.2,
while the measurement design problems are treated as a convex optimization problem as described in Sections 3.3 and 3.4, separately, or with the simultaneous design as proposed in Section 4. As such, there is no simple solution for this multi-factor optimization problem.

In this work, we propose an iterative double-layer procedure, as illustrated in Fig.2, to design the experimental factors for both the input and the observation. The design of input factors is processed in the upper layer, and the integrated observation design is handled in the lower layer.

Due to the non-convex nature of the input design problem, a modern heuristic method - PSO (Kennedy, 2011), is chosen to obtain the optimal solution globally. The PSO method is a population-based optimization algorithm which can solve a variety of hard problems with fast convergent rates. With this algorithm, only a few parameters need to be tuned and no derivative calculations are required, making the algorithm attractive from the computation point of view. The basic PSO method is based on a population of $s$ particles that represent solutions of the optimization problem. Each particle is associated with a position $x$ and a velocity $v$, which denote its position and movement through the searching space. The position and velocity of a particle can be dynamically adjusted via 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 $v_i^{k+1}$:

$$x_i^{k+1} = x_i^k + v_i^{k+1}$$ (21)

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

$$v_i^{k+1} = \omega \cdot v_i^k + \alpha_1 \cdot r_1 \cdot (pbest_i^k - x_i^k) + \alpha_2 \cdot r_2 \cdot (gbest^k - x_i^k)$$ (22)
Figure 2: Iterative double-layer design for both input factors and observation factors
where \( \omega, \alpha_1 \) and \( \alpha_2 \) 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. \( p_{best}^k_i \) is the best position of the \( i \)-th particle at the \( k \)-th generation, and \( g_{best}^k \) is the best position of the \( k \)-th generation among all particles, which can be determined by

\[
g_{best}^k = \arg\min_{z \in x_1^k, x_2^k, \ldots, x_s^k} g(z) \tag{23}
\]

Here \( g(\cdot) \) is the objective function. The pseudo code for PSO implementation is described in Algorithm 5.1 as follows.

**Algorithm 5.1**

1. Choose a population size \( s \) and the iteration number \( n_{tol} \). Initialize the swarm positions \( x_1^0, x_2^0, \ldots, x_s^0 \) and their velocities \( v_1^0, v_2^0, \ldots, v_s^0 \).
2. Let \( p_{best}^0_i = x_i^0, i = 1, 2, \ldots, \) determine \( g_{best}^0 \) using (23), and let \( k = 0 \).
3. Set \( g_{best}^{k+1} = g_{best}^k \). For every particle \( i \), do:
   - Check the constraint of \( x_i^k \), make sure that each particle stays within the bound.
   - If \( f(x_i^k) \leq f(p_{best}^k_i) \), then update the best position of the \( i \)-th particle, \( p_{best}^{k+1}_i = x_i^k \); if \( f(p_{best}^{k+1}_i) \leq f(g_{best}^{k+1}) \), then update the best position at current generation, \( g_{best}^{k+1} = p_{best}^{k+1}_i \); otherwise, set \( p_{best}^{k+1}_i = p_{best}^k_i \).
4. Compute \( x_i^{k+1} \) and \( v_i^{k+1} \) for each particle using equations (21) and (22).
5. Terminate the process when \( k = n_{tol} \). Otherwise, increase \( k \) by one and go to step 3.

With this iterative double-layer structure, the inputs are firstly determined by applying the PSO for a pre-defined number of iterations, based on
which the 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 an up-
dated design of the input factor. This process lasts until the optimal solution
is obtained. While the optimization at the lower layer can solve the convex
optimization problem of observation design under the given input conditions,
the upper-layer design employing stochastic searching largely increases the
chance of finding a global solution for input factors. This is a clear advan-
tage over the traditional local numerical algorithms which most likely only
lead to local optimum. For a complex OED 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 optimization
problem that is relatively easy to solve. The main procedure of the iterative
double-layer optimization design is given in the following.

Algorithm 5.2

1. Initialize the overall OED objective function $g(x, y)$, where $x$ and $y$ de-
note the input and observation variables, respectively. Set the stopping
tolerance level $\delta_{\text{tol}} \geq 0$.
2. Let the iteration number $l = 0$, use the Powell’s method to calculate
$y_{\text{best}}^0$ based on $x_{\text{set}}$. $x_{\text{set}}$ is a vector of pre-setting values for the input
variables. Then determine $x_{\text{best}}^0$ for $g(x, y_{\text{best}}^0)$ using Algorithm 5.1.
3. For iteration $l$, determine $x_{\text{best}}^l$ for the objective function $g(x, y_{\text{best}}^{l-1})$ in
the upper layer using the PSO method described in Algorithm 5.1, then calculate $y_{\text{best}}^l$ for $g(x_{\text{best}}^l, y)$ in the lower layer using the Powell’s
method.
4. If \(|g^{l+1} - g^l| \leq \delta_{tol}\), then stop the optimization process. Otherwise, increase \(l\) by one and go back to step 3.

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

6. Case studies on two biochemical reaction processes

6.1. Enzymatic process with kinetically controlled synthesis reactions

The first case study system is an enzymatic process with kinetically controlled synthesis reactions as illustrated in Fig. 3. In this reaction system, \(S\) is the donor substrate, \(P\) is the leaving group product, \(N\) denotes the nucleophile, \(Q\) is the desired product, \(R\) is the hydrolysis by-product; \(W\) stands for water whose quantity is taken as constant due to its large amount; \(E\) is the enzyme and \(ES\), \(E^*\), \(EQ\) and \(ER\) are different complex forms of enzymes. All reactants are assumed to be well mixed in the reactor. At the beginning of the reactions, the initial reactant species are the donor substrate, the nucleophile and the catalyst. The substrate firstly binds to the enzyme 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 to form \(EQ\) or be hydrolyzed to produce \(ER\). The compound \(EQ\) can be decomposed into the required product and enzyme,
while $ER$ can be decomposed into the hydrolysis by-product and enzyme. During the whole reaction process, all the reactions are reversible, except for the decomposition of $ER$ to give $E$ and $R$. Due to the characteristics of enzyme, it only catalyzes the reactions and at the end of the reaction, the amount of enzyme remains the same as before the chemical reactions.

![Figure 3: Enzyme kinetically controlled synthesis process](image)

A number of enzymatic processes have the similar kinetically controlled synthesis reaction scheme, e.g., in the preparation of semi-synthesis penicillins, $S$ is hydroxyphenylglycine methyl ester and $N$ is 6-APA, etc. In this system, the desired product $Q$ is not thermodynamically the most favourable one. The hydrolysis by-product $R$ will dominate at long times. Among those reaction species, $Q$, $S$, $P$, $N$ and $R$ are measurable in experiments while it is difficult to measure different forms of enzymes due to its very low concentrations. The initial concentrations of $S$, $N$ and $E$ are user-controllable inputs written as $S_0$, $N_0$ and $E_0$, respectively. It is known that a chem-
Figure 4: Time profiles of the 5 measurable state variables of the enzyme reaction system.

The enzyme reaction is always affected by the surrounding environment and some other factors such as the inactivation of enzyme, reactant instability, effects of pH and temperature, etc. In order to investigate the system with a focus on experimental design, all these complications have been removed in this test system. Following the mass balance principle, the enzyme reaction system can be expressed as 10 ODEs including 11 parameters, as given in Appendix B. The nominal values of the model parameters and the initial conditions of the state variables are listed in Table B.5. The time profiles of the 5 measurable state variables under the nominal model parameters and initial conditions are illustrated in Fig. 4. More modeling details and system analysis of this enzyme reaction system can be found in (Yue et al., 2013).

6.1.1. Structural identifiability analysis

To determine whether the model parameters are structurally identifiable, the generating series approach combined with the identifiability tableau, as
introduced in Section 2.2, are implemented to the enzyme reaction model.
Ideally it is always possible to obtain a full rank Jacobian matrix for the power
series coefficients because the number of Lie derivatives of model equations
is infinite.

Through the numerical steps as proposed in (Balsa-Canto et al., 2010),
the Jacobian matrix of the series coefficients with respect to model parame-
ters can be obtained and shown in the tableaus shown in Fig. 5. 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 non-zero element with
respect to the model parameter in the corresponding column. In Fig. 5(a),
there are 27 non-zero series coefficients with respect to the 11 model parame-
ters, which are obtained by the Lie derivative computations. Fig. 5(b) shows
a reduced tableau where 11 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, iter-
atively, until the tableau cannot be further reduced. For the enzyme reaction
system, the final minimum tableau is shown in Fig. 5(c), in which only five
parameters are remained that need to be further checked to see whether they
are structurally identifiable or not. When all the five measurable state vari-
ables are included in the observation, all of the 11 model parameters can
be determined as globally structurally identifiable. When a subset of the
five states are included in the observation, some parameters are found to be locally structurally identifiable or even not identifiable.

Figure 5: Identifiability tableaus based on generating series approach

6.1.2. Sensitivity analysis and parameter estimability analysis

Linear correlation between parameter pairs will affect parameter estimability. The correlation coefficient between two parameters, $k_i$ and $k_j$, can be calculated from FIM as $R_{ij} = \frac{\text{cov}(k_i, k_j)}{\sqrt{\text{FIM}_{ii} \times \text{FIM}_{jj}}}$. Parameters $k_i$ and $k_j$ are said to be linearly correlated if $R_{ij} = 1$ or $-1$. In the correlation matrix composed of $R_{ij}$, any non-diagonal entries with values close to $\pm 1$ suggests a strong correlation between the pair of parameters. Simulation results show high correlations between several reversible reaction pairs in this system although they are not fully correlated. The orthogonalized sensitivity analysis (Yao et al., 2003) is implemented (see the algorithm in Appendix A) instead of the standard local sensitivity analysis (LSA). The ranking of parameters in terms of their influence to the states and the correlation coefficients between parameter pairs are shown in Fig. 6. The metric of 'IEOS' in Fig. 6 represents the integrated effect of the model parameters to the model outputs by using the orthogonalized sensitivity analysis.
Considering both local sensitivities and correlations between all parameters, three parameters, \( k_2, k_{-3} \) and \( k_5W \) are found to be the most important and most identifiable among the 11 parameters. This result is largely consistent with our previous analysis based purely on LSA, where \( k_2, k_{-3} \) and \( k_{-5} \) were identified to be the top three most important parameters (Yue et al., 2013). With the orthogonalization-based method employed in this work, parameter \( k_5W \) replaces \( k_{-5} \) in the top 3 key parameters, which may due to the fact that \((k_5W, k_{-5})\) is a highly correlated parameter pair. In the following OED simulation studies on this case study enzyme reaction system, these three key parameters are considered in the design.

![Parameter ranking](image1.jpg)  
(a) Parameter ranking  

![Parameter correlations](image2.jpg)  
(b) Parameter correlations  

Figure 6: Orthogonalization-based sensitivity analysis & correlation analysis results

### 6.1.3. Observation design and results

We start from observation design by taking the nominal parameter values and the initial conditions in Appendix B. The design objectives are: (i) to select measurement state variables; and (ii) to locate 100 measurement
data points for the selected state variables, which will lead to the most in-
formative data set for parameter estimation. In the simulation, 10% relative
measurement errors and 0.001 absolute measurement errors are added to the
simulation data. Three different design strategies, as shown in Table 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 optimization problem and the solutions
for both can be obtained simultaneously. The D-optimal design criterion is
employed in all the three OED methods.

<table>
<thead>
<tr>
<th>OED methods</th>
<th>Design procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy 1</td>
<td>Sequential: measurement set → sampling time profile</td>
</tr>
<tr>
<td>Strategy 2</td>
<td>Sequential: sampling time profile → measurement set</td>
</tr>
<tr>
<td>Strategy 3</td>
<td>Simultaneous: measurement set &amp; sampling profile</td>
</tr>
</tbody>
</table>

The design results of the three observation strategies and also the default
setting without any OED are listed in Table 2. When no OED is employed,
all the 5 measurable states are taken into account, and the same uniform
sampling rule is applied to all the 5 states, i.e., 20 sampling points for each
state. 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 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.

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 state variables, the sampling points distribution from different experimental strategies are shown in Fig. 7. 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 designed for S and two for Q, respectively, covering a wide 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 confidence interval (CI) of the three key parameters in pairs are com-
pared in Fig. 8. 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 better result 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 best result among the three methods due to the fact that all observation factors are considered simultaneously during the OED. The OED of observation provides a useful insight that the sampling points should be taken at certain regions that corresponds to large parameter sensitivities or large change rates in key variables, not necessarily equally spaced as in a traditional way.

Table 2: Observation design results for enzyme reaction system

<table>
<thead>
<tr>
<th>Methods</th>
<th>Selected measurement states</th>
<th>Sampling time points (unit: second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no-OED</td>
<td>{S, P, N, Q, R}</td>
<td>[300:300:6000]</td>
</tr>
<tr>
<td>Strategy 1</td>
<td>{S, Q}</td>
<td>[450:30:870], [2670:30:3120], [3390:30:4530]</td>
</tr>
<tr>
<td>Strategy 2</td>
<td>{S, P, N, Q, R}</td>
<td>[510:30:690], [2790:30:2940], [4200:30:4380]</td>
</tr>
<tr>
<td>Strategy 3</td>
<td>{S}</td>
<td>[420:30:1020], [2130:30:3390]</td>
</tr>
<tr>
<td></td>
<td>{Q}</td>
<td>[30:30:240], [3930:30:4740]</td>
</tr>
</tbody>
</table>

Note: each region of the sampling time profile, in all tables, is shown as [initial time: sample interval: final time]
Figure 7: Sampling profiles of S and Q with different OED strategies.
Figure 8: Comparison of confidence intervals with different observation design (enzyme reaction system)
6.1.4. Iterative two-layer design of input and observation

In this section, input and observation variables are designed together through the proposed iterative double-layer OED strategy as shown in Fig. 2. The results are compared to another iterative OED, but the observations are designed using Strategy 1, as discussed in Section 6.1.3. In both cases, the iteration number is set to be 100. In both methods, the typical run time of the optimization process is around 1.5 hours on a personal computer with i5-2400 CPU and 4GB memory. The designed results are shown in Table 3. By considering both the input and observation factors, S and Q are both selected for the measurement set, which is consistent with the observation design results in Section 6.1.3. 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 Fig. 9. Again, it can be seen that using the same computational time, the results from the proposed OED method provides (potentially) better parameter estimation quality compared with the method with sequential observation design.

One should note that using the proposed observation design or the iterative double-layer design strategy, non-uniform sampling time regions are suggested to do the measurement rather than the uniform sampling strategy. The latter has been widely used in chemical engineering. Taking uniform sampling at the very early stage of modeling and design would be useful, where model information is limited and parameter values contain large uncertainties.
Table 3: Iterative two-layer OED of input and observation

<table>
<thead>
<tr>
<th>Two-layer OED</th>
<th>$[S_0, E_0, N_0]$: (unit: mol/L)</th>
<th>measur. set</th>
<th>Sampling points (unit: second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower layer:</td>
<td>$[0.74, 1.52e-5, 1]$</td>
<td>{S,Q}</td>
<td>$[420:30:1020]$,</td>
</tr>
<tr>
<td>Strategy 1</td>
<td>$[2130:30:3390]$ for S; and</td>
<td></td>
<td>$[30:30:240]$,</td>
</tr>
<tr>
<td></td>
<td>$[3930:30:4740]$ for Q</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower layer:</td>
<td>$[1, 5.64e-6, 0.13]$</td>
<td>{S,Q,N}</td>
<td>$[4590:30:5760]$ for S; and</td>
</tr>
<tr>
<td>Strategy 3</td>
<td>$[5280:30:6000]$ for Q; and</td>
<td></td>
<td>$[390:30:1410]$ for N</td>
</tr>
<tr>
<td></td>
<td>$[390:30:1410]$ for N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.2. Enzymatic biodiesel production system

A kinetic model for a lab-scale enzymatic transesterification of rapeseed oil with methanol using Callera Trans L (a liquid formulation of a modified *Thermomyces lanuginosus* lipase) was developed in (Price et al., 2014). In this model, the methanol inhibition and the interfacial and bulk concentrations of the enzyme are considered except for the enzyme deactivation process. The developed model describes the effect of different oil compositions, as well as different water, enzyme and methanol concentrations, which are the relevant conditions required for process evaluation of industrial production of biodiesel. Fig. 10 demonstrates the reaction scheme of this enzymatic 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. Then these enzyme
substrates can be decomposed into the acyl enzyme complex and D, M and G, respectively. The acyl enzyme complex can then 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. From these kinetic reactions a set of ODEs can be formulated following the mass-balance principle (Appendix C).

A set of experiments have been conducted in advance in order to collect
experimental data for parameter estimation. In all those experiments, the contents of water and enzyme were varied from 3 to 7 and 0.1 to 0.5 wt.% oil respectively. An amount of 1.5 equivalents of methanol was reacted with the Rapeseed oil. One equivalent corresponds to the stoichiometric amount of alcohol needed to convert all fatty acid residues in the oil to biodiesel. The reaction was carried out in a 0.25 liter glass reactor with a tank diameter (T) of 55 mm and 2 baffles, each is 0.18T wide. The reactor was immersed in a water bath with temperature control maintained at 35°C. Initially 0.2 equivalent of methanol was charged with the oil in the reactor. When the reaction mixture reached the reaction temperature, the amount of water and enzyme to be used in the experiment was then added to the reactor and methanol feeding started. The experiment length is set to be 25 hours and
original samplings take place every 15 minutes in the first hour and then once each hour. The unit for all reactant concentrations is in mol/L. The nominal parameter values, initial conditions and feeding rates are provided in Table C.7 and C.8 in Appendix C.

The orthogonalization-based method is applied to rank parameters and examine parameter correlations so as to select the set of most estimable parameters. This method gives consistent results regarding the 10 estimable parameters using the collinearity index. The three most important parameters identified in this analysis are $k_6$, $k_8$ and $k_9$ (shown in Fig. 11).

Figure 11: Parameter ranking via orthogonalization (enzymatic biodiesel production system)

Taking the three most important parameters, $k_6$, $k_8$ and $k_9$, into the parameter estimation scheme, OED has been applied to determine the best observation strategy which include the most valuable measurement variables and the best sampling time points for each state. Considering the reality
of experimentation, the minimal sampling time interval between two neighboring sampling points is set to be 5 minutes. In non-designed settings, 28 equally spaced sampling points were selected for all five measurable state variables which are T, D, M, BD and FFA. The number of sampling points in this simulation is therefore chosen to be 140 (28 × 5). Three different experimental strategies in Table 1 are tested, the results of which are shown in Table 4 and in Fig. 12.

<table>
<thead>
<tr>
<th>Table 4: Design results of different OED strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measurement state variables</strong></td>
</tr>
<tr>
<td>no-OED</td>
</tr>
<tr>
<td>Strategy 1</td>
</tr>
<tr>
<td>Strategy 2</td>
</tr>
<tr>
<td>Strategy 3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

All three OEDs give two sampling regions on the selected or all measurement variables. The measurement should be taken for BD and FFA at the start (first 200 minutes) and middle (between 600 and 1000 minutes) stages of the reaction. This is reasonable because the changes of FFA and M are significant from the start of the reaction. Sampling points selected in this region can grab dynamic information of the system. Also, from Fig. 12(d) it can be seen that the sensitivities of key parameters to BD in the middle reaction stage are quite high. Therefore, additional samplings should be taken in
this region. Of course the sampling details are not the same when different
OEDs are implemented. The sequential designs with Strategy 1 and Strategy
2 show that BD and FFA are the most valuable state variables, while in the
integrated observation design of Strategy 3, sampling points for M are also
shown to be useful. The observation design results are further assessed by
comparing the CIs of key parameter pairs in Fig. 13. 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.
Figure 12: Sampling points on selected state variables (enzymatic biodiesel production system)
Figure 13: Comparison of confidence intervals for different OED strategies (enzymatic biodiesel production system)
7. Conclusions and discussions

In this work, three experimental design objectives, the input design, the sampling time design and the measurement set selection are investigated. The integrated observation design that determines both measurement set selection and sampling time scheduling, simultaneously, has been proposed. By approximating available sampling points \textit{a priori}, the problem formulation for sampling time design can be expressed in a similar form of measurement set selection design. Therefore these two design tasks can be combined together as a single optimization problem, which is further relaxed to a convex optimization problem that can be conveniently solved using local optimization methods. Furthermore, we have developed an iterative two-layer numerical strategy which can deal with OED taking into account input and observation variables together. This new optimization strategy intends to obtain the global optimal results for all experimental conditions in one optimization framework. The input design that is formulated as non-convex optimization problem is solved by a modern heuristic algorithm, PSO method, in the upper layer. The integrated observation design which can be relaxed into convex optimization problem is solved by a local optimization method, the Powell’s method, in the lower layer. In each iteration, the local optimization and the global optimization are handled separately.

Through the case studies based on an enzyme reaction model and a kinetic model developed for a lab-scale enzyme-catalysed biodiesel production process, the effectiveness of the integrated observation design over two traditional sequential design strategies has been examined. In both case studies, the lower bounds for parameter estimation errors can be reduced through the
proposed 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. Similar improvement can be observed when the input is included in the iterative two-layer design. The resulted non-uniform sampling time selection is rather non-intuitive but could be of more values compared with conventional uniform sampling schedule. Whether the non-uniform sampling schedule is generally appropriate for wider applications need more investigations in future work. It is expected that a well-designed sampling schedule contains more useful information and a non-uniform sampling should be more cost effective compared with conventional uniform sampling.

OED is a model-based technology, the results of which depend on the prior knowledge of the system model, also on the design criteria and the optimization methods used. It is therefore not always possible to get consistent OED results under various circumstances. Nevertheless, this is a systematic method that can provide useful guidance to experimental settings, with the benefits of collecting measurement data that are most valuable to process modeling. The OED results can sometimes be different from the experiences or intuitive understanding of the experimental conditions, for example, key regions in sampling rather than uniform sampling for measurement data can be revealed by OED, which may not be obvious from the experimental practice. The measurement set selection may suggest useful variables that are ignored in the existing measurement system. Further development on model-based OED methodology can be investigated by considering more
complicated factors required by applications, for example, model uncertainties during the design stage, non-Gaussian noise in measurement, design of time-dependent experimental factors. All these tasks will be very challenging.

Acknowledgment

The authors would like to thank Dr Jason Price and his colleagues from the Department of Chemical and Biochemical Engineering, Technical University of Denmark, for providing the mathematical model of the enzymatic biodiesel production system and for many useful discussions.

Bibliography


design for parameter estimation in unstructured growth models. Biotechn-
ology Progress 10 (5), 480–488.

optimization of bioprocesses: Efficient and robust numerical strategies.

timal identification experiments for nonlinear dynamic process models: a
stochastic global optimization approach. Industrial & Engineering Chem-
istry Research 41 (10), 2425–2430.

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

neous strategies for dynamic process optimization. Chemical Engineering
Science 57 (4), 575–593.

Optimum design of experiments for enzyme inhibition kinetic models. Jour-

Brown, M., He, F., Yeung, L. F., 2008. Robust measurement selection for
biochemical pathway experimental design. International Journal of Bioin-


Appendix A. Orthogonalized sensitivity analysis

The basic step of orthogonalization based forward selection method is described as follows.

1. The normalized parameter sensitivity $\bar{S}$ and the magnitude of each column in $\bar{S}$ is calculated based on (13) and (3), respectively. The parameter corresponding to the column with maximum magnitude is the first identifiable parameter. Set $k=1$.

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

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

\[
\hat{\bar{S}}_k = X_k (X_k^T \cdot X_k)^{-1} \cdot X_k \bar{S}
\]

and calculate the residual matrix by $R_k = \bar{S} - \hat{\bar{S}}_k$.

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 $\bar{S}$ that corresponds to the $(k+1)$-th parameter to matrix $X_k$.
6. Repeat steps 3-5 for all parameters until the maximum magnitude in $\mathbf{R}_k$ is less than a predefined threshold.

Appendix B. Supplementary materials of the enzyme reaction system

The 10 ordinary differential equations for the enzyme reaction system are as follows.

\[
\begin{align*}
\frac{dE}{dt} &= -k_1 \cdot E \cdot S + k_{-1} \cdot ES + k_4 \cdot EQ - k_{-4} \cdot E \cdot Q \\
&\quad + k_6 \cdot ER \quad \text{(B.1)} \\
\frac{dES}{dt} &= k_1 \cdot E \cdot S - k_{-1} \cdot ES - k_2 \cdot ES + k_{-2} \cdot E^* \cdot P \quad \text{(B.2)} \\
\frac{dE^*}{dt} &= k_2 \cdot ES - k_{-2} \cdot E^* \cdot P - k_3 \cdot E^* \cdot N + k_{-3} \cdot EQ \\
&\quad - k_5 \cdot W \cdot E^* + k_{-5} \cdot ER \quad \text{(B.3)} \\
\frac{dEQ}{dt} &= k_3 \cdot E^* \cdot N - k_{-3} \cdot EQ - k_4 \cdot EQ + k_{-4} \cdot E \cdot Q \quad \text{(B.4)} \\
\frac{dER}{dt} &= k_5 \cdot W \cdot E^* - k_{-5} \cdot ER - k_6 \cdot ER \quad \text{(B.5)} \\
\frac{dS}{dt} &= -K_1 \cdot E \cdot S + k_{-1} \cdot ES \quad \text{(B.6)} \\
\frac{dP}{dt} &= k_2 \cdot ES - k_{-2} \cdot E^* \cdot P \quad \text{(B.7)} \\
\frac{dN}{dt} &= -k_3 \cdot E^* \cdot N + k_{-3} \cdot EQ \quad \text{(B.8)} \\
\frac{dQ}{dt} &= k_4 \cdot EQ - k_{-4} \cdot E \cdot Q \quad \text{(B.9)} \\
\frac{dR}{dt} &= k_6 \cdot ER \quad \text{(B.10)}
\end{align*}
\]
Table B.5: List of state variables and kinetic parameters

<table>
<thead>
<tr>
<th>State variables</th>
<th>Initial condition</th>
<th>Kinetic parameters</th>
<th>Nominal values</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S(x_1) )</td>
<td>0.8</td>
<td>( k_1 )</td>
<td>1e5</td>
</tr>
<tr>
<td>( P(x_2) )</td>
<td>0</td>
<td>( k_{-1} )</td>
<td>1e3</td>
</tr>
<tr>
<td>( N(x_3) )</td>
<td>0.9</td>
<td>( k_2 )</td>
<td>100</td>
</tr>
<tr>
<td>( Q(x_4) )</td>
<td>0</td>
<td>( k_{-2} )</td>
<td>1e4</td>
</tr>
<tr>
<td>( R(x_5) )</td>
<td>0</td>
<td>( k_3 )</td>
<td>5e4</td>
</tr>
<tr>
<td>( E(x_6) )</td>
<td>1.5e-5</td>
<td>( k_{-3} )</td>
<td>200</td>
</tr>
<tr>
<td>( E^*(x_7) )</td>
<td>0</td>
<td>( k_4 )</td>
<td>1e3</td>
</tr>
<tr>
<td>( ES(x_8) )</td>
<td>0</td>
<td>( k_{-4} )</td>
<td>2e4</td>
</tr>
<tr>
<td>( EQ(x_9) )</td>
<td>0</td>
<td>( k_5 ) ( W )</td>
<td>5e3</td>
</tr>
<tr>
<td>( ER(x_{10}) )</td>
<td>0</td>
<td>( k_{-5} )</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( k_6 )</td>
<td>500</td>
</tr>
</tbody>
</table>
Figure B.14: Parameter relative sensitivities to S, N, P, R (enzyme reaction system)
Figure B.15: Parameter relative sensitivities to non-measurable enzyme complexes (enzyme reaction system)
Figure B.16: Parameter ranking for enzyme reaction system
Appendix C. Supplementary information of enzymatic biodiesel production system

Ordinary differential equations of enzymatic biodiesel production system:

\[
\begin{align*}
\frac{d([T] \cdot V)}{dt} &= -V(r_2) \\
\frac{d([D] \cdot V)}{dt} &= V(r_3 - r_4) \\
\frac{d([M] \cdot V)}{dt} &= V(r_5 - r_6) \\
\frac{d([BD] \cdot V)}{dt} &= V(r_9) \\
\frac{d([FA] \cdot V)}{dt} &= V(r_8) \\
\frac{d([G] \cdot V)}{dt} &= V(r_7) \\
\frac{d([W] \cdot V)}{dt} &= -V(r_8) \\
\frac{d([CH] \cdot V)}{dt} &= -V(r_9 + r_{10}) \\
\frac{d([E] \cdot V)}{dt} &= V(r_1 + r_8 + r_9 - r_2 - r_4 - r_6 - r_{10}) \\
\frac{d([EX] \cdot V)}{dt} &= V(r_3 + r_5 + r_7 - r_8 - r_9) \\
\frac{d([ET] \cdot V)}{dt} &= V(r_2 - r_3) \\
\frac{d([ED] \cdot V)}{dt} &= V(r_5 - r_6) \\
\frac{d([EM] \cdot V)}{dt} &= V(r_6 - r_7) \\
\frac{d([ECH] \cdot V)}{dt} &= V(r_{10}) \\
\frac{d([E_{bulk}] \cdot V)}{dt} &= -V(r_1) \\
\frac{d(V_p)}{dt} &= R_G + R_W \\
\frac{d(V)}{dt} &= (F_a)
\end{align*}
\]
Table C.6: Kinetic mechanism for the enzymatic transesterification

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Rate Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{bulk} + A_f \leftrightarrow E$</td>
<td>$r_1 = k_1 \cdot E_{bulk} \cdot A_f - k_{-1} \cdot E$</td>
</tr>
<tr>
<td>$T + E \leftrightarrow ET$</td>
<td>$r_2 = k_2 \cdot T \cdot E - k_{-2} \cdot ET$</td>
</tr>
<tr>
<td>$ET \leftrightarrow EX + D$</td>
<td>$r_3 = k_3 \cdot ET - k_{-3} \cdot EX \cdot D$</td>
</tr>
<tr>
<td>$D + E \leftrightarrow ED$</td>
<td>$r_4 = k_4 \cdot D \cdot E - k_{-4} \cdot ED$</td>
</tr>
<tr>
<td>$ED \leftrightarrow EX + M$</td>
<td>$r_5 = k_5 \cdot ED - k_{-5} \cdot EX \cdot M$</td>
</tr>
<tr>
<td>$M + E \leftrightarrow EM$</td>
<td>$r_6 = k_6 \cdot M \cdot E - k_{-6} \cdot EM$</td>
</tr>
<tr>
<td>$EM \leftrightarrow EX + G$</td>
<td>$r_7 = k_7 \cdot EM - k_{-7} \cdot EX \cdot G$</td>
</tr>
<tr>
<td>$EX + W \leftrightarrow FA + E$</td>
<td>$r_8 = k_8 \cdot EX \cdot W - k_{-8} \cdot FA \cdot E$</td>
</tr>
<tr>
<td>$EX + CH \leftrightarrow BD + E$</td>
<td>$r_9 = k_9 \cdot EX \cdot CH - k_{-9} \cdot BD \cdot E$</td>
</tr>
<tr>
<td>$CH + E \leftrightarrow ECH$</td>
<td>$r_{10} = k_{10} \cdot CH \cdot E - k_{-10} \cdot ECH$</td>
</tr>
</tbody>
</table>
Table C.7: Nominal parameter values for enzyme biodiesel production system

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>4.95e4</td>
</tr>
<tr>
<td>$k_{-1}$</td>
<td>6.60</td>
</tr>
<tr>
<td>$k_2$</td>
<td>1.69e6</td>
</tr>
<tr>
<td>$k_{-2}$</td>
<td>1.11e4</td>
</tr>
<tr>
<td>$k_3$</td>
<td>2.07e4</td>
</tr>
<tr>
<td>$k_{-3}$</td>
<td>2.20e7</td>
</tr>
<tr>
<td>$k_4$</td>
<td>3.41e6</td>
</tr>
<tr>
<td>$k_{-4}$</td>
<td>1.33e7</td>
</tr>
<tr>
<td>$k_5$</td>
<td>1.55e7</td>
</tr>
<tr>
<td>$k_{-5}$</td>
<td>1.81e5</td>
</tr>
<tr>
<td>$k_6$</td>
<td>9.13e4</td>
</tr>
<tr>
<td>$k_{-6}$</td>
<td>5.43e5</td>
</tr>
<tr>
<td>$k_7$</td>
<td>7.06e6</td>
</tr>
<tr>
<td>$k_{-7}$</td>
<td>4.93</td>
</tr>
<tr>
<td>$k_8$</td>
<td>2.36e4</td>
</tr>
<tr>
<td>$k_{-8}$</td>
<td>3.51e6</td>
</tr>
<tr>
<td>$k_9$</td>
<td>2.54e4</td>
</tr>
<tr>
<td>$k_{-9}$</td>
<td>2.05e5</td>
</tr>
<tr>
<td>$k_{10}$</td>
<td>3.23e-2</td>
</tr>
<tr>
<td>$k_{-10}$</td>
<td>4.39e-4</td>
</tr>
</tbody>
</table>

Figure C.17: Time profile of state variables for enzymatic biodiesel production system
Table C.8: Initial input values and feeding rate of methanol

<table>
<thead>
<tr>
<th>Species</th>
<th>Ini. cond. $(mol \cdot L^{-1})$</th>
<th>Species</th>
<th>Ini. cond. $(mol \cdot L^{-1})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T(x_1)$</td>
<td>0.9536</td>
<td>$EX(x_{10})$</td>
<td>0</td>
</tr>
<tr>
<td>$D(x_2)$</td>
<td>0.0195</td>
<td>$ET(x_{11})$</td>
<td>0</td>
</tr>
<tr>
<td>$M(x_3)$</td>
<td>0.0014</td>
<td>$ED(x_{12})$</td>
<td>0</td>
</tr>
<tr>
<td>$B(x_4)$</td>
<td>1e-4</td>
<td>$EM(x_{13})$</td>
<td>0</td>
</tr>
<tr>
<td>$FFA(x_5)$</td>
<td>0.0224</td>
<td>$ECH(x_{14})$</td>
<td>0</td>
</tr>
<tr>
<td>$G(x_6)$</td>
<td>1e-6</td>
<td>$Ef(x_{15})$</td>
<td>9.7165e-6</td>
</tr>
<tr>
<td>$W(x_7)$</td>
<td>2.3854</td>
<td>$Vp(x_{16})$</td>
<td>0.0661</td>
</tr>
<tr>
<td>$CH(x_8)$</td>
<td>0.5850</td>
<td>$V(x_{17})$</td>
<td>1.5383</td>
</tr>
<tr>
<td>$E(x_9)$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Methanol feed rate [eq $\cdot h^{-1}$] | Initial dose methanol [eq] | water [wt.% oil] | Enzyme [wt.% oil] |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.185 first 2hrs; 0.06 thereafter</td>
<td>0.2</td>
<td>5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

64