Sequential Model-Based A- and V-Optimal Design of Experiments for Building Fundamental Models of Pharmaceutical Production Processes

Ali Shahmohammadi , Kimberley B. McAuley

PII: S0098-1354(19)30417-X
DOI: <https://doi.org/10.1016/j.compchemeng.2019.06.029>
Reference: CACE 6504

To appear in: *Computers and Chemical Engineering*

Received date: 16 April 2019
Revised date: 25 June 2019
Accepted date: 26 June 2019

Please cite this article as: Ali Shahmohammadi , Kimberley B. McAuley , Sequential Model-Based A- and V-Optimal Design of Experiments for Building Fundamental Models of Pharmaceutical Production Processes, *Computers and Chemical Engineering* (2019), doi: <https://doi.org/10.1016/j.compchemeng.2019.06.029>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights

- Sequential model-based design of experiments results in informative experimental data for estimating parameters and making predictions using fundamental models.
- Inversion of Fisher Information Matrix (**FIM**) is required for model-based design of experiments.
- The FIM may be noninvertible for complex chemical and pharmacological process models.
- Parameter subset selection leaves out problematic parameters, resulting in a reduced **FIM** that is effective for experimental design.

Sequential Model-Based A- and V-Optimal Design of Experiments for Building Fundamental Models of Pharmaceutical Production Processes

Ali Shahmohammadi; Kimberley B. McAuley *

Department of Chemical Engineering, Queen's University, Kingston, ON K7L 3N6, Canada

AUTHOR INFORMATION

Corresponding Author

* Kimberley B. McAuley

Phone: (613) 533-6000 ext. 77973

E-mail: kim.mcauley@queensu.ca

Co-Author

Ali Shahmohammadi

E-mail: ali.shahmohammadi@queensu.ca

ABSTRACT

Sequential model-based A- and V-optimal experimental designs are known to be effective for maximizing the information content of data, leading to reliable parameter estimates and model predictions. A- and V-optimal designs require inversion of the Fisher Information Matrix (FIM), which may be noninvertible especially for fundamental models with many parameters. In this study, two different methodologies for selecting sequential approximately A- and V-optimal experiments are compared for situations where the FIM is noninvertible. The first approach, called Leave Out (LO) approach, finds and leaves out problematic parameters that make the FIM noninvertible and the second approach, called Pseudoinverse (PI) approach, uses a Moore-Penrose pseudoinverse of the FIM. Comparisons are carried out using a Michaelis Menten reaction dynamic model for production of a pharmaceutical agent. Monte Carlo simulations indicate, for both A- and V-optimal situations, that designed experiments using the LO approach are superior to designs obtained by the PI approach.

Keywords: Sequential design of experiments; A-optimal; V-optimal; Noninvertible Fisher information matrix; Leave Out approach; Pseudoinverse approach.

1 INTRODUCTION

Application of mathematical models for chemical and pharmacological processes is widespread since they can provide valuable information for process design and simulation (Pandey and Bharadwaj, 2016). Especially in pharmaceutical production processes, mathematical models offer crucial information at multiple stages in the drug development and scale-up the process, thereby saving time and effort. Empirical models are commonly used to aid the development of pharmaceutical production processes by providing equations that link experimental inputs to the outputs of interest (Benson, 2018; Boukouvala et al., 2011; Domagalski et al., 2015; Hallow et al., 2010; Huang et al., 2009). The main disadvantage of the empirical models is that they are usually not suitable for obtaining predictions for dynamic process behavior nor for predicting behavior outside the range of experimental conditions used in model development. For this reason, mechanistic (fundamental) models developed based on the physics and chemistry of the system are more desired (Birtwistle et al., 2013; Domagalski et al., 2015).

Fundamental models play a key part in the Quality by Design (QbD) paradigm (Hallow et al., 2010; Huang et al., 2009; Jain, 2014; Yu et al., 2014). QbD requires systematic incorporation of scientific knowledge and risk assessment principles to develop pharmaceutical products that lie within the acceptable predefined range required for patient safety (FDA, 2004; Jain, 2014; Yu et al., 2014). Mechanistic models are also used by pharmaceutical companies to develop continuous manufacturing processes, which are replacing traditional batch processes (Azzaro-Pantel, 2018; Benyahia et al., 2012; Yu et al., 2014). Having predictive models that can relate process inputs (i.e., raw material attributes) to the model outputs (i.e., product attributes) is advantageous because models can be used for valuable quantitative risk assessments are multiple

stages of the drug production process (Azzaro-Pantel, 2018). Application of fundamental models for a variety of pharmaceutical processes has been discussed by several researchers (Benyahia et al., 2012; Gernaey and Gani, 2010; Hallow et al., 2010; Petrides et al., 2010, 2002; Poós and Szabó, 2017). For example, Hallow et al., (2010) developed a fundamental model to describe lactamization and ethanolysis reactions in a pharmaceutical production process (Hallow et al., 2010). Benyahia et al., (2012) developed a plant-wide fundamental model for a continuous pharmaceutical manufacturing process with several unit operations (Benyahia et al., 2012). Poós and Szabó (2017) developed a fundamental model to investigate the effects of operating parameters that influence simultaneous heat and mass transfer in a fluidized bed dryer (Poós and Szabó, 2017).

Fundamental models of chemical or biochemical processes often contain many unknown parameters that require estimation using informative experimental data (Maria, 2004). It is advantageous to carefully select experimental settings using a formal experimental design technique prior to performing new experiments (Atkinson, 1982). A variety of model-based design of experiment (MBDoE) techniques has been developed to assist modelers in selecting experiments that maximize the information content of the data (Box et al., 2005). Most common MBDoE techniques (e.g., A-, D- and E- optimal designs) have been developed to improve the quality of the parameter estimates (Franceschini and Macchietto, 2008), while other techniques (e.g., G- and V- optimal designs) assist modelers in improving the quality of model predictions at desired operating conditions (Liu and Neudecker, 1995). There are also MBDoE techniques (e.g., T-optimal design) aimed at model discrimination (i.e., to select the best model among several rival models) (Atkinson and Fedorov, 1975). Using MBDoE techniques is advantageous because they take into account the structure of the model as well as parameters and measurement

uncertainties when selecting new experiments (John and Draper, 1975; Welch, 1984). Using MBDoE, it is straightforward to select any number of experiments that match available resources for experimentation (Quaglio et al., 2018).

Conventional MBDoE techniques involve solution of nonlinear constrained optimization problems to minimize uncertainties in parameter estimates or model predictions (Franceschini and Macchietto, 2008). As indicated by the objective functions in Table 1, these MBDoE calculations rely on the inverse of the Fisher Information Matrix (**FIM**) (Shirt et al., 1994; Walter and Pronzato, 1990). The **FIM** plays a significant role in MBDoE and parameter estimation problems because it carries information about how the model predictions are affected by changes in parameter values (Walter and Pronzato, 1990; Zhang and Edgar, 2008). For models that are linear in the parameters, the **FIM** is independent of the parameter values. However, for nonlinear models (which are commonly used to describe chemical and biochemical systems), the **FIM** is computed based on linearization of the model around its parameter values (Atkinson, 1982; Box et al., 2005; Box and Lucas, 2006; Ford et al., 1989; Pinto et al., 1990). Since the true parameter values are unknown, estimated or assumed parameter values are used. If parameter values used for linearization differ too much from their true values, the selected experimental settings obtained using MBDoE might be ineffective (Bauer et al., 2000). To avoid this problem, experiments can be designed and performed sequentially to update parameter values and experimental protocols (e.g., changes in the bounds of the design space) as new experimental information becomes available (Ford and Silvey, 1980). Sequential MBDoE is also appealing because it provides an opportunity to incorporate information from any old experimental data that may have been collected for other purposes. These old data may contain

limited, but helpful, information to aid parameter estimation and design of new experiments (Bauer et al., 2000; Issanchou et al., 2005).

Table 1. Optimality criteria for model-based design of experiments (Atkinson and Fedorov, 1975; Dette and Wong, 1999; Franceschini and Macchietto, 2008; Liu and Neudecker, 1995; Shirt et al., 1994; Walter and Pronzato, 1990)

Optimality Criterion	Description	Eq. #
$J_A = \text{trace}((\mathbf{FIM})^{-1})$	A-optimal design minimizes total parameter variance.	(1.1)
$J_D = \det((\mathbf{FIM})^{-1})$	D-optimal design minimizes the volume of the joint confidence interval for the parameters.	(1.2)
$J_E = \lambda_{\max}(\mathbf{FIM}^{-1})$	E-optimal design minimizes the largest eigenvalue of the FIM , thereby minimizing the uncertainties in the worst-case direction in the parameter space.	(1.3)
$J_G = \max(\text{diag}(\mathbf{W}(\mathbf{FIM})^{-1}\mathbf{W}^T))$	G-optimal design minimizes the maximum variance of model predictions at user-specified operating conditions of interest, specified using a matrix W . This is equivalent to minimizing the largest value of the diagonal elements of $\mathbf{W}(\mathbf{FIM})^{-1}\mathbf{W}^T$.	(1.4)
$J_T = -\text{trace}((\mathbf{FIM})^{-1})$	T-optimal is used to discriminate between two rival models by treating one model as the true model. The selected experiments maximize the sum of squares of lack of fit for the second model.	(1.5)
$J_V = \text{trace}(\mathbf{W}(\mathbf{FIM})^{-1}\mathbf{W}^T)$	V-optimal design minimizes the total variance of model predictions at user-specified operating conditions of interest, which are specified using matrix W . This is equivalent to minimizing the trace of $\mathbf{W}(\mathbf{FIM})^{-1}\mathbf{W}^T$.	(1.6)

Model users are often interested in using mathematical models to predict system responses over a region of interest (Welch, 1984). As shown in Table 1, G- and V-optimal designs focus on selecting experiments for obtaining accurate predictions at operating condition

specified by the modeler. These important settings, where reliable predictions are desired, are specified using the matrix $\mathbf{W}$. G-optimal design minimizes the *maximum* variance of the model predictions and V-optimal design minimizes the *total* variance of the model predictions at the settings of interest (Liu and Neudecker, 1995; Welch, 1984). In this way, G-optimal design selects experiments to improve the worst-case model predictions at conditions specified by the matrix $\mathbf{W}$. By contrast, V-optimal design selects experiments to improve the model predictions for all settings of interest in the $\mathbf{W}$ matrix (Welch, 1984). There are relatively few applications of G- and V-optimal designs in the literature (François et al., 2004; Goos and Syafitri, 2014; Liu and Neudecker, 1995; Stigler, 1971; Wong, 1994). Wong (1994) compared G-optimal design with A-, D- and E-optimal designs for a variety of polynomial models (Wong, 1994). Liu and Neudecker (1995) and Goos & Syafitri (2014) designed V-optimal experiments for predicting properties of mixtures of several components (Goos and Syafitri, 2014; Liu and Neudecker, 1995). François et al. (2004) used G- and V- optimal designs for a univariate nonlinear calibration model (François et al., 2004). Sequential model-based V-optimal design was used by Thompson et. al, (2010) to obtain accurate model prediction for the molecular-weight-distribution of ethylene copolymers produced using Ziegler-Natta-catalyzed (Thompson et al., 2010). In the current article, V-optimal design is considered for selecting experimental settings in a pharmaceutical production process where there is a desire to reduce the variance of model predictions at settings of interest. To reduce parameter variance, we will also consider A-optimal designs, which have been more widely used (Barz et al., 2013; Bauer et al., 2000; López C. et al., 2015; López C et al., 2013; Shit et al., 1994; Walter and Pronzato, 1990).

Obtaining A- and V-optimal experimental settings, using objective functions (1.1) and (1.6), requires an invertible **FIM**. However, for complex chemical and biochemical models with

many unknown parameters, the **FIM** may not be invertible (Barz et al., 2013; Ben-Zvi et al., 2004; Cui et al., 2015; Issanchou et al., 2003; Karimi et al., 2012; Kou et al., 2005a; Littlejohns et al., 2010; López C. et al., 2015; López C et al., 2013; Murphy et al., 2004; Nguyen et al., 2017; Thompson et al., 2010; Yue et al., 2006; Zhao et al., 2017). A singular FIM causes problems in both designing experiments and parameter estimation (S. Vajda et al., 1989). Reasons for a noninvertible **FIM** are problems with i) structural identifiability and ii) practical identifiability. A structural identifiability problem arises when the model is not well-structured so that unique estimates for the parameters cannot be obtained, even using a large quantity of noise-free data with well-designed experiments (Banga and Balsa-Canto, 2008; Ljung and Glad, 1994). Structural identifiability issues can be addressed by obtaining new types of observations and by updating the model structure to include these observed responses. Also, the model can be reformulated or reparameterized to remove the non-identifiable parameters (Asprey and Naka, 1999; Espie and Macchietto, 1988; McLean and McAuley, 2012; Quaglio et al., 2019). Several different approaches have been developed to investigate structural identifiability problems including: a Taylor-series-expansion approach (Pohjanpalo, 1978), a generating-series approach (Walter and Lecourtier, 1982), a similarity transformation approach (Sandor Vajda et al., 1989) and a differential algebra approach (Ben-Zvi et al., 2010; Ljung and Glad, 1994). Practical identifiability (also known as estimability) problems result from using insufficient experimental data during parameter estimation (McLean and McAuley, 2012; Petersen et al., 2001). A simple approach for investigating estimability is to check if the **FIM** is of full rank (Petersen et al., 2001). If the rank of the **FIM** is lower than the number of parameters, the parameters cannot be estimated uniquely using the available experimental data (Holmberg, 1982). Conducting more

experiments may resolve estimability problems (McLean and McAuley, 2012). Note that the current article focuses mainly on problems of practical identifiability (estimability).

A noninvertible **FIM** arises in many chemical, biochemical and pharmacological systems because experimental data may be limited and the models are often complex with a large number of unknown parameters (e.g., 10-80 parameters) (Balsa-Canto et al., 2010; Ben-Zvi et al., 2004; Butcher, 2005; Cho et al., 2003; Cui et al., 2015; DiMasi et al., 2003; Issanchou et al., 2003; Jain, 2014; Kou et al., 2005a; Littlejohns et al., 2010; Murphy et al., 2004; Nguyen et al., 2017; Thompson et al., 2010; Yue et al., 2006; Zhao et al., 2017). Several studies have addressed the problem of a noninvertible **FIM** when designing experiments using MBDoE (Barz et al., 2013; López C. et al., 2015; López C et al., 2013; Srinath and Gunawan, 2010; Thompson et al., 2010; Yue et al., 2006). Common approaches include parameter subset selection, e.g. (Chu et al., 2009; Kravaris et al., 2013; Wu et al., 2008), pseudoinverse methods, e.g. (Eghtesadi and McAuley, 2016; Greville, 1959; Li and Yeh, 2012), Tikhonov regularization, e.g. (Hoerl and Kennard, 1970; Johansen, 1997; Tibshirani, 2011; Tikhonov et al., 1978) and Bayesian approaches, e.g. (Chaloner and Verdinelli, 1995; Durán and White, 1995; Ruggoo and Vandebroek, 2004). In a recent study, we considered the problem of a noninvertible **FIM** when selecting A-optimal settings using several case studies including a nonlinear dynamic model of a pharmaceutical production process (Shahmohammadi and McAuley, 2019). We compared the effectiveness of two approaches: i) use of a Moore-Penrose pseudoinverse (PI) of the **FIM** in place of $\mathbf{FIM}^{-1}$ when optimizing objective function in equation (1.1) and ii) use of a parameter-subset selection technique to find problematic parameters, which are left out of the optimization, resulting in a reduced **FIM**, $\mathbf{FIM}_r$. These case study results suggest that A-optimal MBDoE using $\mathbf{FIM}_r$ leads to better parameter estimates than the Moore Penrose PI. One shortcoming of our previous study

is that the investigations were carried out on a simple models (i.e., linear models and a relatively simple nonlinear example). Further investigation on more complicated nonlinear models is required to validate the results.

To our knowledge, there are no studies that address the problem of noninvertible **FIM** when selecting sequential V-optimal settings. An investigation into how to handle a singular **FIM** during V-optimal design is conducted in the current study. We are also interested to find out if it is better to fix problematic parameters (LO approach) at their nominal values to avoid creating an ill-conditioned problem or if it is better to use Pseudoinverse (PI approach) to address ill-conditioning during sequential MBDOE. Therefore, the two main objectives of the current study are i) to use a more complicated pharmaceutical model to test conclusions from our previous study on how to handle noninvertible **FIMs** during A-optimal design and ii) to investigate the effectiveness of PI and LO methodologies for selecting sequential V-optimal settings when the **FIM** is noninvertible. In addition, we compare the resulting A-optimal and V-optimal experimental designs to explore the benefits of V-optimal designs when accurate predictions are desired. The fundamental dynamic model investigated in this study is based on the system described by Domagalski et al. (2015) (Domagalski et al., 2015), which uses Michaelis–Menten kinetics and enzyme-catalyzed reactions (Lehninger, 2009) of interest to our industrial sponsor.

First, background information regarding the FIM and sequential A- and V-optimal design are presented. Next, details of the parameter-subset selection and Moore Penrose PI are provided, along with the procedure used for comparing the LO and PI approaches. The pharmaceutical case study and the corresponding dynamic model are then presented. Results obtained using Monte Carlo (MC) simulations are provided, revealing that the LO approach is superior to the PI

approach when the **FIM** is noninvertible, for both A- and V-optimal designs. Finally, results for A-optimal and V-optimal designs are compared with results from a more traditional approach where the modeler selects conditions from the corners of the permissible operating space.

2 BACKGROUND INFORMATION

2.1 Fisher Information Matrix for Nonlinear Models

Consider a nonlinear model of the form:

$$\mathbf{Y} = \mathbf{g}(\mathbf{d}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{Y} \in \mathbf{R}^N$ is a vector of stacked measured responses obtained from experimentation, $\mathbf{g}$ is the solution of a set of algebraic or differential equations that describe the behavior of the experiments, $\mathbf{d} \in \mathbf{R}^{r \times D}$ is a matrix of experimental settings (for r runs with D decision variables specified for each). $\boldsymbol{\theta} \in \mathbf{R}^p$ is the vector of model parameters and $\boldsymbol{\varepsilon} \in \mathbf{R}^N$ is the vector of random measurement noise, which is assumed to have a diagonal covariance matrix $\boldsymbol{\Sigma} \in \mathbf{R}^{N \times N}$. Note that, for dynamic multi-response models with n sampling times per run and v response variables, the total number of data values is $N = nvr$. A complete list of symbols and acronyms is provided in the nomenclature section at the end of the manuscript.

Computation of the **FIM** for nonlinear models in the form of equation (1) depends on a parametric sensitivity matrix $\mathbf{S} \in \mathbf{R}^{N \times p}$. The elements of $\mathbf{S}$:

$$S_{ij} = \left. \frac{\partial g(\mathbf{d}, \boldsymbol{\theta})}{\partial \theta_j} \right|_{\theta_k \neq j} \quad (2)$$

matrix may be obtained by linearizing the model around its initial parameter guesses $\boldsymbol{\theta}_0 \in \mathbf{R}^p$. Elements of $\mathbf{S}$ are scaled using measurement uncertainties and parameter uncertainties to form a scaled local sensitivity matrix $\mathbf{Z}$ with elements (Thompson et al., 2009):

$$Z_{ij} = S_{ij} \frac{s_{\theta_j}}{s_{y_i}} \quad (3)$$

The scaling factor s_{y_i} is a user-specified uncertainty for the i th measurement, which might be obtained from replicated experiments. Similarly, s_{θ_j} is a user-specified uncertainty for the initial guess of the j th parameter. based on the modeler's knowledge or assumption about what values might be reasonable (Thompson et al., 2010). As such, a small value of s_{θ_j} indicates that the modeler is confident about the initial guess for the j th parameter. Scaling the elements of the sensitivity matrix is advantageous because it accounts for the modeler's level of confidence about measurements and initial parameter guesses and it makes the elements of the sensitivity matrix dimensionless (Thompson et al., 2009). The **FIM** can be computed from the scaled sensitivity matrix:

$$\mathbf{FIM} = \mathbf{Z}^T \mathbf{Z} \quad (4)$$

When designing sequential model-based optimal experiments, $\mathbf{Z}$ contains two parts:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_{\text{old}} \\ \mathbf{Z}_{\text{new}} \end{bmatrix} \quad (5)$$

where $\mathbf{Z}_{\text{old}}$ corresponds to experimental settings and data from old experiments. The elements of $\mathbf{Z}_{\text{new}}$ depend on the settings for new sequential experiments that will be designed. As a result, $\mathbf{Z}_{\text{old}}$ (which contains $N_{\text{old}} = n_{\text{old}} v_{\text{old}} r_{\text{old}}$ rows and p columns) is fixed during sequential MBDoE and elements of $\mathbf{Z}_{\text{new}}$ are determined based on the new experimental settings obtained by optimizer.

The $\mathbf{W}$ matrix for V-optimal design is a scaled sensitivity matrix obtained based on operating the conditions of interest to the modeler. The elements of $\mathbf{W}$ are:

$$W_{ij} = \left. \frac{\partial g(\mathbf{d}_w, \boldsymbol{\theta})}{\partial \theta_j} \frac{s_{y_i}}{s_{\theta_j}} \right|_{\theta_{k \neq j}} \quad (6)$$

where the $\mathbf{d}_w \in \mathbf{R}^{r_w \times D}$ is the matrix of operating conditions of interest for r_w sets of operating conditions at which accurate predictions are desired.

To obtain a sequential A- or V-optimal design, the optimizer selects values of the operating conditions that minimize $J_A = \text{trace}(\mathbf{FIM}^{-1}) = \text{trace}((\mathbf{Z}^T \mathbf{Z})^{-1})$ or $J_V = \text{trace}(\mathbf{W}(\mathbf{FIM}^{-1})\mathbf{W}^T) = \text{trace}(\mathbf{W}(\mathbf{Z}^T \mathbf{Z})^{-1}\mathbf{W}^T)$. When designing experiments using several sequential in steps, the sensitivity matrix $\mathbf{Z}_{\text{old}}$ is updated after each round of experiments to include additional rows resulting from the most recent experiments. It is also common to update $\mathbf{Z}_{\text{old}}$ by re-linearizing around the improved parameter values obtained using the full sets of experiments to date.

2.2 Approaches for dealing with a noninvertible FIM for parameter estimation

When the **FIM** is noninvertible only a subset of the model parameters can be estimated uniquely (Eghtesadi and McAuley, 2014; Kravaris et al., 2013; López C et al., 2013; Yao et al., 2003). In this situation, the modeler may find problematic parameters that lead to noninvertibility and fix them at their nominal values, so that the remaining (nonproblematic) parameters can be estimated. The problem of finding the subset of parameters that can be and should be estimated has long been of interest to many research groups (Burth, 1999; Chu et al., 2009; Eghtesadi et al., 2013; Lund and Foss, 2008; Mallows, 1973; Neshat and Pope, 2017; Velez-Reyes and Verghese, 1995; Yao et al., 2003). An orthogonalization-based ranking approach (see Table 2) developed by Yao et al. and Thompson et al., (Thompson et al., 2009; Yao et al., 2003) has been used by many researchers e.g., (Casas-Orozco et al., 2018; Chu and Hahn, 2009; Jayasankar et al., 2009; Karimi et al., 2017, 2012; Kou et al., 2005b; Kravaris et al., 2013; Meshram et al., 2013; Ngo et al., 2013). This orthogonalization approach ranks parameters from most-estimable to least-estimable to find the problematic parameters that make the **FIM** noninvertible.

Steps used in the orthogonalization-based parameter ranking methodology are shown in Table 2. In step 1, the magnitude (i.e., sum of squares) of each column of $\mathbf{Z}$ is computed. In step

2, the parameter that corresponds to the column with the largest magnitude is selected as the most-estimable parameter. This column is put into the matrix $\mathbf{X}_k$ (i.e., $k = 1$ for the first iteration). In step 3, using equation (2.1), the columns of the $\mathbf{Z}$ matrix are regressed onto columns of $\mathbf{X}_k$. Residual matrix $\mathbf{R}_k$ is computed using equation (2.2) to remove correlation between unranked parameters, and parameters that have been already ranked. In step 4, magnitudes of the columns of $\mathbf{R}_k$ are computed and the parameter corresponding to the column with the largest magnitude is selected as the next-most-estimable parameters. In step 5, $\mathbf{X}_k$ is augmented by including the column from $\mathbf{Z}$ corresponding to this parameter, resulting in $\mathbf{X}_{k+1}$. Steps two to five are continued for $k = 2, 3, \dots, p$ to produce a ranked list of up to p parameters. The ranking procedure stops after the iteration when $\mathbf{X}_k^T \mathbf{X}_k$ (i.e., the reduced **FIM**) becomes noninvertible. Problematic parameters (those that are unranked when the invertibility problem occurs) tend to be parameters that have small influence on model predictions of the available (or proposed) data, have effects that are correlated with those of other parameters, or have values that are well known by the modeler (i.e., parameters with relatively small s_{θ_j}). Unranked parameters can be fixed at their nominal values or can be removed from the model via simplifying assumptions (McLean and McAuley, 2012; Thompson et al., 2009).

The orthogonalization-based ranking approach is helpful for avoiding numerical problems during parameter estimation. Nevertheless, estimating all of the ranked parameters may lead to overfitting of the available data (Chu et al., 2009; Wu et al., 2011, 2008). When a model suffers from overfitting, it may give an acceptable fit to the data but could lead to poor predictions for new settings that are not used during parameter estimation (Chu et al., 2009). To avoid overfitting and to obtain accurate model predictions, Wu et al.(2011) developed a mean-squared-error (MSE) parameter subset selection criterion (Wu et al., 2011). Tables S1 and S2

(Supplementary information) show the detailed procedure for computing Wu's critical ratio, r_{cc} , and using it (along with the orthogonalization algorithm in Table 2) to select a subset of parameters that should be estimated. The parameter subset with the smallest value of the r_{cc} is the one that is expected to give predictions with the smallest MSE (at conditions where experimental results are available). The methodology in Tables S1 and S2 (Supplementary information) has been used for parameter subset selection and estimation in variety of models (e.g., Cui et al., 2015; Woloszyn et al., 2013; Karimi, Schaffer and McAuley, 2012; Cui *et al.*, 2013; Zhao *et al.*, 2013) and will be used as a part of current study.

Table 2. Orthogonalization algorithm (Thompson et al., 2009; Yao et al., 2003)

1. For each column of scaled sensitivity matrix $\mathbf{Z}$, compute the sum of squares (i.e., the Euclidean norm). Select the column with the largest magnitude as the most estimable parameter. Set $k = 1$.
2. Construct the matrix $\mathbf{X}_k$ by including the k selected columns from $\mathbf{Z}$ that correspond to parameters that have been ranked.
3. Use $\mathbf{X}_k$ to predict columns in $\mathbf{Z}$ using ordinary least squares:

$$\hat{\mathbf{Z}}_k = (\mathbf{X}_k^T \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{Z} \quad (2.1)$$

and calculate the residual matrix:

$$\mathbf{R}_k = \mathbf{Z} - \hat{\mathbf{Z}}_k \quad (2.2)$$

4. Calculate the magnitude of each column in $\mathbf{R}_k$. The $(k + 1)^{\text{th}}$ -most estimable parameter corresponds to the column in $\mathbf{R}_k$ with the largest magnitude.
5. Put the columns corresponding to the $k + 1$ parameters that have been ranked in the matrix $\mathbf{X}_k$.
6. Advance the iteration counter k by one and repeat Steps 2–5, until all parameters are ranked or until it is impossible to perform the least-squares prediction of $\mathbf{Z}$ in step 3 due to matrix singularity.

A common alternative to parameter subset selection for avoiding a noninvertible **FIM** during parameter estimation is to use a pseudoinverse (PI) of the **FIM**. The Moore Penrose PI, $\mathbf{FIM}^{*-1}$, which is the most popular PI in the statistics literature (e.g., Greville, 1959; Rao, 1966; Nayak and Foudriat, 1974; Eldén, 1982; Li and Yeh, 2012) is computed using:

$$\mathbf{FIM} = \mathbf{USV} = [\mathbf{u}_1 \ \mathbf{u}_2] \begin{bmatrix} \mathbf{s}_1 & 0 \\ 0 & 0 \end{bmatrix} [\mathbf{v}_1 \ \mathbf{v}_2] \quad (7)$$

$$\mathbf{FIM}^{*-1} = \mathbf{u}_1 \mathbf{s}_1^{-1} \mathbf{v}_1 \quad (8)$$

where $\mathbf{u}_1$ and $\mathbf{v}_1$ correspond to nonzero singular values (i.e., $\mathbf{s}_1$) of the **FIM** and $\mathbf{u}_2$ and $\mathbf{v}_2$ correspond to zero or very small (smaller than a predefined threshold) singular values of the **FIM**. Other methods to address a noninvertible/singular **FIM** include using Bayesian approaches e.g., (Chaloner and Verdinelli, 1995) or Tikhonov regularization approaches (e.g., LASSO estimation or ridge regression (Hoerl and Kennard, 1970; Tibshirani, 2011)).

In the current case-study paper, the methodology shown in Table 2 is adapted for use in sequential A-and V-optimal designs, to deal with the noninvertible **FIM**. As in our recent work, this approach is called the LO approach, and the effectiveness of the LO approach for MBDoE is compared with a Moore-Penrose-based PI approach. Often modelers might be interested in estimating the full sets of parameters, however, due to estimability (and possibly structural identifiability) problems, it might be impossible to estimate all of the parameters using the available dataset. In this situation, modelers might be interested in getting more informative data. The LO and PI approaches can help with sequential MBDoE to obtain data that will lead to better parameter estimates and model predictions even if the full set of parameters can not be estimated uniquely using the new data. After obtaining new data and estimating a large subset of parameters, modelers may decide to fix the remaining parameters at their nominal values or to remove them by simplifying the model.

3 PROCEDURE FOR COMPARING THE LO AND PI APPROACHES

Table 3 summarizes how MC simulations are used in this study to compare the LO and PI approaches. The first step is to generate simulated old experimental data. Next, a subset of the parameters is estimated from the old data. In step 3, initial guesses for the decision variables and

the parameter estimates from step 2 are used to form an initial $\mathbf{Z}_{\text{new}}$ matrix, $\mathbf{Z}_{\text{new}0}$, and $\mathbf{Z}_0 = \begin{bmatrix} \mathbf{Z}_{\text{old}} \\ \mathbf{Z}_{\text{new}0} \end{bmatrix}$. In step 4, the LO approach is used to design sequential A- or V-optimal experiments based on the reduced $\mathbf{FIM}$, $\mathbf{FIM}_r$. Note that for V-optimal design a reduced $\mathbf{W}$ matrix, $\mathbf{W}_r$, is also constructed based on the ranked (nonproblematic) parameters. In step 5, the PI approach is used to select approximate A- or V-optimal settings. In step 6, simulated data are generated for the settings obtained from the LO and PI approaches. In step 7, parameters are estimated using the simulated data (old and new) for both approaches and SSD_θ and SSD_{Y_W} are computed for each approach using equations (3.1) and (3.2). Steps 1 to 7 are repeated 100 times (i.e., starting from 100 different old simulated datasets) so that boxplots can be constructed for the values of SSD_θ and SSD_{Y_W} , to compare the effectiveness of the LO and PI approaches. The reason for repeating the procedure 100 times is that we found 100 times is sufficient to show clear patterns in the behavior of the two approaches so that informed decisions can be made. A larger number of MC simulations could be conducted but would require additional computational expense, especially for the nonlinear parameter estimations.

SSD_θ in equation (3.1) is a measure of closeness of parameter estimates to their corresponding true values (used to generate the simulated data). This criterion is related to the A-optimality criterion and, as such, is useful for comparing the effectiveness of the LO and PI approaches for A-optimal design when the FIM is noninvertible. Similarly, SSD_{Y_W} is a measure of how close the model predictions are to the noise-free system outputs at the conditions of interest. This criterion is related to the V-optimality criterion and, as such, is useful for comparing the effectiveness of the LO and PI approaches for V-optimal design. Note that, in real experimental design situations, true parameter values are unknown, and the modeler designs new sequential experiments only once. The MC simulation results in the current article are

valuable because they shed light on which approach, i.e., LO or PI, would be better on average for selecting new experimental conditions.

Table 3. Procedure for comparing the effectiveness of LO and PI approaches for designing sequential A- and V-optimal experiments

1. Use MC simulations to generate simulated data for old experimental settings (i.e., settings corresponding to $\mathbf{Z}_{old}$), using the true parameter values $\boldsymbol{\theta}^{true}$ and noise variance σ_i^2 .
2. Estimate parameters from the old synthetic data using the orthogonalization + r_{cc} method in Table S2.
3. Use the parameter estimates from step 2 and initial guesses for the decision variables to construct $\mathbf{Z}_{new0}$ as well as $\mathbf{Z}_0 = \begin{bmatrix} \mathbf{Z}_{old} \\ \mathbf{Z}_{new0} \end{bmatrix}$, where $\mathbf{Z}_{old}$.
4. *LO method:* Rank parameters using the orthogonalization method in Table 2 and $\mathbf{Z}_0$. Ranking stops when $\mathbf{X}_k^T \mathbf{X}_k$ (in step 2 of Table 2) becomes noninvertible (i.e., condition number > a user-specified cut-off value). Classify the remaining unranked parameters as problematic parameters and fix them at their initial values. Form the reduced sensitivity matrix $\mathbf{Z}_r$ using the ranked parameters and compute $\mathbf{FIM}_r = \mathbf{Z}_r^T \mathbf{Z}_r$.
 - *A-optimal design:* Use $\mathbf{FIM}_r$ in the objective function (1.1) in Table 1 to select approximately A-optimal values of the decision variables.
 - *V-optimal design:* Form the reduced sensitivity matrix $\mathbf{W}_r$ for the conditions of interests. Use $\mathbf{FIM}_r$ and $\mathbf{W}_r$ in objective functions (1.6) in Table 1 to select approximate V-optimal values of the decision variables.

Using the updated $\mathbf{Z}$ matrix, re-rank the parameters and update the list of problematic parameters. If the list of left-out parameter changes, update $\mathbf{Z}_r$, $\mathbf{FIM}_r^{-1}$ (and $\mathbf{W}_r$ if applicable) and repeat the optimization. Continue until the list of left-out parameters does not change.

5. *PI method:* Use $\mathbf{FIM}^{*-1}$ in place of $\mathbf{FIM}^{-1}$ in objective functions for A- or V-optimal design to determine optimal values of the decision variables. A user-specified cut-off for the condition number is used to determine the number of singular values considered in $\mathbf{FIM}^{*-1}$.
6. *For each approach (i.e., LO and PI):* Use true values of the parameters $\boldsymbol{\theta}^{true}$ and noise variances σ_i^2 to generate simulated data for the proposed new experiments.
7. Estimate parameters from available data (old and new) and compute: SSD_{θ} and SSD_{Y_W}

$$SSD_{\theta} = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{true})^T \boldsymbol{\Sigma}_{\theta}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{true}) \quad (3.1)$$

$$SSD_{Y_W} = (\mathbf{g}(\mathbf{d}_w, \hat{\boldsymbol{\theta}}) - \mathbf{g}(\mathbf{d}_w, \boldsymbol{\theta}^{true}))^T \boldsymbol{\Sigma}^{-1} (\mathbf{g}(\mathbf{d}_w, \hat{\boldsymbol{\theta}}) - \mathbf{g}(\mathbf{d}_w, \boldsymbol{\theta}^{true})) \quad (3.2)$$

where $\hat{\boldsymbol{\theta}} \in R^p$ is the vector of estimated parameters (including the fixed values of any parameters that were held at initial guesses) and $\boldsymbol{\theta}^{true}$ is vector of true parameter values. $\boldsymbol{\Sigma}_{\theta} \in R^{p \times p}$ is a diagonal matrix containing squared values of the user provided parameter uncertainties s_{θ_j} .

8. Repeat steps 1-7 for 100 MC simulations

4 MICHAELIS–MENTEN CASE STUDY

4.1 Reaction scheme and dynamic model

The nonlinear dynamic batch-reactor model considered in the current study involves Michaelis–Menten kinetics for production of a pharmaceutical agent. The reaction scheme in Figure 1 was used by Domagalski et al. (2015) to build an empirical model using conventional DoE and response surface methodology (Domagalski et al., 2015). As shown in Figure 1, the reaction scheme starts with a reversible reaction between reagent SM1 and catalyst D to generate intermediate $SM1 \cdot D$. Next, intermediate $SM1 \cdot D$ reacts with reagent SM2 to form the product P, releasing the catalyst D. The reaction scheme includes formation of four impurities: product P and SM2 may react to form impurity I1, SM1 may hydrolyze due to the presence of water to form I2, catalyst D may deactivate with water to form impurity I3, and P may thermally degrade to form impurity I4. Simulated data generated by Domagalski et al., are consistent with the fundamental model equations in Table 4, which describe the time evolution of the reactants and product concentrations in a batch reactor. Equations (4-11) to (4-14) indicate that the concentrations of SM1, D, SM2, and P are measured, and these measurements have experimental errors.

Figure 1. Reaction scheme for Michaelis–Menten case study

Table 4. Dynamic kinetic model

$$\frac{dC_{SM1}}{dt} = -k_{1f}C_{SM1}C_D + \frac{k_{1f}}{K_{eq}}C_{SM1D} - k_4C_{SM1}C_{H_2O} \quad (4-1)$$

$$\frac{dC_D}{dt} = -k_{1f}C_{SM1}C_D + \frac{k_{1f}}{K_{eq}}C_{SM1D} + k_2C_{SM2}C_{SM1D} - k_5C_D C_{H_2O} \quad (4-2)$$

$$\frac{dC_{SM1D}}{dt} = k_{1f}C_{SM1}C_D - \frac{k_{1f}}{K_{eq}}C_{SM1D} - k_2C_{SM2}C_{SM1D} \quad (4-3)$$

$$\frac{dC_{SM2}}{dt} = -k_2C_{SM2}C_{SM1D} - k_3C_{SM2}C_P \quad (4-4)$$

$$\frac{dC_P}{dt} = k_2C_{SM2}C_{SM1D} - k_3C_{SM2}C_P - k_6C_P \quad (4-5)$$

$$\frac{dC_{H_2O}}{dt} = -k_4C_{SM1}C_{H_2O} - k_5C_D C_{H_2O} \quad (4-6)$$

$$\frac{dC_{I1}}{dt} = k_3C_{SM2}C_P \quad (4-7)$$

$$\frac{dC_{I2}}{dt} = k_4C_{SM1}C_{H_2O} \quad (4-8)$$

$$\frac{dC_{I3}}{dt} = k_5C_D C_{H_2O} \quad (4-9)$$

$$\frac{dC_{I4}}{dt} = k_6C_P \quad (4-10)$$

$$y_{SM1} = C_{SM1} + \varepsilon_{SM1} \quad (4-11)$$

$$y_D = C_D + \varepsilon_D \quad (4-12)$$

$$y_{SM2} = C_{SM2} + \varepsilon_{SM2} \quad (4-13)$$

$$y_P = C_P + \varepsilon_P$$

(4-14)

Domagalski et al. (2015) reported a simulated dataset resulting from 60 different experimental settings selected using three rounds of virtual experimentation (i.e., 16 fractional factorial experiments and 4 center-point runs in each round of experimentation). Table 5 provides the conditions for the center-point experimental run in their first round of experimentation. We assume that data from four old replicate runs at these conditions are available to the modeler, so these center-point settings are used to construct $\mathbf{Z}_{old}$.

To reduce the complexity of the modeling problem, it is assumed that the volume of the solution and the water content are constant at $V = 1.0$ L and $C_{H_2O} = 0.10$ M, respectively. The duration of the simulated batch experiment is 6.0 h with measurements taken every 45 minutes, resulting in 9 sampling times, including the initial time $t = 0$. As a result, each run involves $4 \times 9 = 36$ measured values (i.e., 9 values each for y_{SM1} , y_D , y_{SM2} and y_P).

Table 5. Initial conditions for center-point batch reactor operation at $T = 40$ °C

State variable	Units	Initial condition
C_{SM1}	M	1
C_D	M	0.05
C_{SM2}	M	1.15
C_{SM1D}	M	0
C_P	M	0
C_{H_2O}	M	0.10
C_{I1}	M	0
C_{I2}	M	0
C_{I3}	M	0
C_{I4}	M	0

Domagalski et al., (2015) used assumed true values of the kinetic coefficients (see Table 6) to generate simulated data with independent normally-distributed random errors. Values in Table 6 can be used to compute kinetic and equilibrium coefficients at different temperatures from the Arrhenius expressions:

$$k_i(T) = k_{i,ref} \exp\left(-\frac{E_{a,i}}{R}\left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right) \quad (9)$$

$$K_{eq_1}(T) = K_{eq_1,ref} \exp\left(-\frac{\Delta H_1}{R}\left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right) \quad (10)$$

where k_i is the i th kinetic coefficient, R is universal gas constant, T is the temperature in K, and $T_{ref}=313.15$ K = 40 °C is a reference temperature. In equation (10), K_{eq_1} is the equilibrium kinetic coefficient for the reaction 1, and ΔH_1 is corresponding reaction enthalpy.

Table 6. True values of the kinetic coefficients and equilibrium constant (Domagalski et al., 2015)

	Units	Value at 40 °C	$E_{a,i}$ or ΔH_1 (J. mol ⁻¹)
k_{1f}	M ⁻¹ s ⁻¹	1.09×10^{-01}	5.00×10^4
K_{eq}	M ⁻¹	9.33	-1.00×10^4
k_2	M ⁻¹ s ⁻¹	3.39×10^{-03}	4.25×10^4
k_3	M ⁻¹ s ⁻¹	1.09×10^{-06}	9.00×10^4
k_4	M ⁻¹ s ⁻¹	1.17×10^{-05}	9.50×10^4
k_5	M ⁻¹ s ⁻¹	1.93×10^{-06}	9.75×10^4
k_6	s ⁻¹	1.87×10^{-08}	8.00×10^4

In the current study, parameter values in Table 6 and measurement noise variances in Table 7 are used to generate simulated datasets. Measurement variances in Table 7 are consistent with the variability of the center-point runs reported by Domagalski et al. (2015). Figure 2 compares simulation results obtained using the true parameter values in Table 6 with simulated data obtained at 40 °C. The noise levels ($\sigma_{SM1}^2 = 4.8 \times 10^{-2}$ M², $\sigma_D^2 = 2.1 \times 10^{-4}$ M², $\sigma_{SM2}^2 = 5.2 \times 10^{-2}$ M², $\sigma_P^2 = 5.0 \times 10^{-2}$ M²) for generating simulated data points in Figure 2 are consistent with those shown in Table 7. As shown in Figure 2b, the concentration of the catalyst drops rapidly to ≈ 0.007 M, due to reaction (1f). Two different scenarios are considered

in this case study: i) operation at $T = T_{ref} = 40\text{ }^{\circ}\text{C}$, and ii) operation at a variety of different temperatures between $35\text{ }^{\circ}\text{C}$ and $45\text{ }^{\circ}\text{C}$. It is assumed that the temperature is controlled accurately at the desired set points. Design of A- and V-optimal experiments for both scenarios is described below, with further details provided in the supplementary information.

Table 7. Measurement variances used for generating simulated data

Measured response (M)	$\sigma_{y_i}^2$ (M ²)
y_{SM1}	4.8×10^{-2}
y_D	2.1×10^{-4}
y_{SM2}	5.2×10^{-2}
y_P	5.0×10^{-2}

Figure 2. Simulated data (symbols) and noise-free concentration profiles (curves) obtained at 40 °C using experimental settings in Table 5, kinetic coefficients in Table 6 and measurement errors in Table 7. i.e., a) SM1; b) D; c) SM2; d) P.

Scenario I. In the first scenario, the parameter vector contains six forward kinetic coefficients and K_{eq1} (i.e., $\theta = [k_{1f}, K_{eq}, k_2, k_3, k_4, k_5, k_6]^T$). In the MBDoe and parameter estimation calculations, the parameter lower bounds are set at zero (i.e., all the kinetic coefficients are positive) and their upper bounds are 5 times their corresponding true values in

Table 6. For designing A- or V-optimal experiments in scenario I, the decision variables are the initial concentrations for the reactants SM1, D and SM2 (i.e., $\mathbf{d} = [C_{SM1_0}, C_{D_0}, C_{SM2_0}]^T$). Table 8 shows lower and upper bounds for these decision variables, which permit experimentation over a wider range of conditions than was considered by Domagalski et al., (2015). Table 9 shows four settings that are assumed to be of particular interest to the modeler, because of the low catalyst levels that are assigned. In the V-optimal design calculations these settings are used to construct $\mathbf{W}$. Note that the lowest catalyst level (0.0091 M) in Table 9 is lower than the lower bound for C_{D_0} in Table 8. This situation is considered in Scenario I because it illustrates the flexible nature of V-optimal designs, wherein the specified operating conditions of interest are permitted to be outside of the range of conditions where the modeler is currently willing to experiment.

Table 8. Lower and upper bounds for the decision variables for the scenario I

Decision variables	Lower bound	Upper bound
M	M	M
C_{SM1_0}	0.10	1.5
C_{D_0}	0.01	0.5
C_{SM2_0}	0.10	1.5

Table 9. Operating conditions of interest where accurate model predictions are desired for scenario I

	C_{SM1_0} (M)	C_{D_0} (M)	C_{SM2_0} (M)
CI_1	1.11	0.011	1.22
CI_2	1.11	0.011	1.33
CI_3	0.91	0.0091	1.22
CI_4	0.91	0.0091	1.33

Scenario II. In the second scenario, temperature is considered as an additional decision variable, i.e., $\mathbf{d} = [C_{SM1_0}, C_{D_0}, C_{SM2_0}, T]^T$. As a result, the parameter vector contains the seven parameters from scenario I, along with ΔH for reaction I and six activation energies, i.e.,

$$\boldsymbol{\theta} = [k_{1f,ref}, K_{eq1,ref}, k_{2,ref}, k_{3,ref}, k_{4,ref}, k_{5,ref}, k_{6,ref}, E_{1f}, \Delta H_1, E_2, E_3, E_4, E_5, E_6]^T. \quad \text{Lower}$$

and upper bounds for all parameters are set at zero and five times the corresponding true values. Lower and upper bounds for decision variables C_{SM1_0} , C_{D_0} , and C_{SM2_0} are the same as in scenario I, with T bounded between 35 °C and 45 °C. Table 10 shows eight different operating conditions (at low catalyst conditions with high and low temperatures) where accurate model predictions are desired. These settings are used to compute **W** for Scenario II.

Table 10. Eight operating conditions of interest for obtaining accurate model predictions in scenario II

	C_{SM1_0} (M)	C_{D_0} (M)	C_{SM2_0} (M)	T (°C)
CI_1	1.11	0.011	1.22	35
CI_2	1.11	0.011	1.33	35
CI_3	0.91	0.0091	1.22	35
CI_4	0.91	0.0091	1.33	35
CI_5	1.11	0.011	1.22	45
CI_6	1.11	0.011	1.33	45
CI_7	0.91	0.0091	1.22	45
CI_8	0.91	0.0091	1.33	45

5 RESULTS AND DISCUSSION

5.1 Scenario I: Operation at 40 °C

Using the old data alone and initial parameter guesses chosen randomly between their lower and upper bounds, only 5 out of 7 parameters could be ranked because the **FIM** was noninvertible. For all initial parameter guesses tested (10 sets of values chosen randomly between lower and upper bounds) parameters k_5 and k_6 were left out of the ranked list using the orthogonalization algorithm in Table 2. Figure 3 shows parameter subsets that were selected using Wu's r_{cc} criterion (see Table S2), and their frequencies for the 100 sets of simulated old experiments. The parameter subset that was selected most often using the old data is $\theta_{sub} = [k_2, K_{eq}]^T$ using 71 out of 100 simulated old datasets. Estimated values of the selected

parameters (along with fixed values of the left-out parameters) were used to compute the initial elements of $\mathbf{Z}$ and $\mathbf{W}$.

Figure 3. Frequency of selected and estimated subset of parameters using r_{cc} criterion for 100 simulated old datasets

Table 11 shows the A-optimal settings that were selected using the LO and PI approaches in 100 MC simulations. Six different sets of conditions were selected using the LO approach and five were selected using the PI approach. In the most frequently selected conditions using the LO approach (LO-A-1), C_{D0} is at its upper bound. In the most frequently selected condition for the PI approach (PI-A-1), both C_{D0} and C_{SM20} are at their upper bounds. To obtain the experimental settings shown in Table 11, the LO approach ranked k_2 as the top parameter followed by K_{eq} most of the time. Using the LO approach, k_6 always (and sometimes k_5 also) was left out and fixed at its nominal value, as shown in Figure S3 in the Supplementary Information. The PI approach fixed three singular values at zero in most cases when selecting new experiments.

Table 11. A-optimal settings for one sequential run selected by the LO and PI approaches

LO approach				
#	C_{SM10} (M)	C_{D0} (M)	C_{SM20} (M)	Frequency from 100 simulations
LO-A-1	0.6773	0.5000	0.1029	60
LO-A -2	0.7726	0.5000	1.5000	17
LO-A -3	1.4838	0.3514	1.5000	13
LO-A -4	1.5000	0.5000	1.0762	6
LO-A -5	1.4969	0.4123	0.1036	2
LO-A -6	0.9222	0.5000	0.3393	2
PI approach				
#	C_{SM10} (M)	C_{D0} (M)	C_{SM20} (M)	Frequency from 100 simulations
PI-A -1	0.7185	0.5000	1.5000	47
PI-A -2	0.6696	0.4997	0.1052	26
PI-A -3	0.9109	0.5000	1.5000	25
PI-A -4	0.8417	0.1937	1.3397	2
PI-A -5	1.1049	0.1460	1.5000	1

New data were generated using the settings in Table 11 and parameters were estimated 100 times using the combined old and new data. Figure 4 shows the subsets of the parameters that were estimated and their corresponding frequencies. The most commonly estimated parameters using experiments designed using the LO approach are k_{1f} and K_{eq1} ; the PI approach resulted in K_{eq1} and k_2 being estimated most often.

Figure 4. Frequency of selected and estimated parameter subsets for 100 (a) LO datasets and (b) PI datasets.

Figure 5 shows corresponding boxplots for SSD_θ (see Table 3) obtained using the various settings in Table 11. The results indicate that the LO approach is superior to the PI approach, on average, for this case study. These results are consistent with our previous study involving a Michael addition reaction with 5 parameters (Shahmohammadi and McAuley, 2019).

Figure 5. Boxplots for 100 values of SSD_θ when designing one sequential A-optimal experiment using LO and PI approaches

The performance of the LO and PI approaches for providing an accurate estimate of individual parameters was also investigated. As shown in Figure 6, both LO and PI approaches resulted in more accurate estimates for k_2 and K_{eq} . The main difference between the results for the two approaches is in estimating k_{1f} , k_3 and k_4 , where the LO approach provides more accurate parameter estimates on average compared to the PI approach. Note that SSD_{θ_i} is computed from:

$$SSD_{\theta_i} = \sum_{l=1}^{100} \left(\frac{\theta_i^{true} - \hat{\theta}_{il}}{s_{\theta_i}} \right)^2 \quad (11)$$

where $\hat{\theta}_{il}$ is the estimate (or corresponding fixed value when the parameter is not estimated) of the i th parameter using the l th MC simulation.

Figure 6. SSD_{θ_i} for each parameter obtained when designing one A-optimal experiments using LO and PI approaches.

Similarly, the LO and PI approaches were also compared for selecting V-optimal experiments. As shown in Figure 7, the LO approach results in model predictions that are closer, on average, to the corresponding noise-free values. Table 12 summarizes the selected settings

for sequential V-optimal designs using LO and PI approaches. The most frequent settings selected by the LO approach has C_{D0} near the lower bound and C_{SM20} at the upper bound. The second most frequent design chosen by the LO approach (LO-V-2) has D_0 and $SM2_0$ at the upper bounds. The PI approach tends to also select C_{D0} and C_{SM20} at the upper bounds. Note that, after designing several additional experiments, all parameters became identifiable and resulted in invertible **FIMs**. Due to invertibility of the **FIM**, the LO and PI approaches resulted in identical settings for the A-optimal designs and identical settings for the V-optimal designs (results not shown).

Table 12. V-optimal settings selected by LO and PI approaches and their frequencies

LO approach				
#	C_{SM10} (M)	C_{D0} (M)	C_{SM20} (M)	Frequency from 100 simulations
LO-V-1	0.7578	0.0797	1.5000	49
LO-V -2	0.7640	0.5000	1.5000	23
LO-V -3	1.5000	0.5000	1.4964	14
LO-V -4	1.1046	0.0979	1.5000	13
LO-V -5	0.9442	0.1313	1.5000	1
PI approach				
#	C_{SM10} (M)	C_{D0} (M)	C_{SM20} (M)	Frequency from 100 simulations
PI-V -1	0.7251	0.5000	1.5000	90
PI-V -2	1.1287	0.2393	1.5000	10

Figure 7. boxplots for 100 values of SSD_{YW} when designing one sequential V-optimal experiment using LO and PI approaches

5.2 Scenario II: operation at a variety temperature between 35 and 45 °C

In this scenario, the LO and PI approaches were used to simultaneously select *two* new sequential A- and V- optimal experiments for a situation where the temperature effect was considered, resulting in a more complicated model with 14 parameters. Similar to scenario I, 100 sets of simulated old data were generated, and parameters were estimated for use in A- and V-optimal designs. Using the old experimental settings alone, k_6 , ΔH and all of the activation of energies were left out of the ranked list, as expected, because there was no temperature variation in the old data. From the ranked list only k_3 was estimated in 73 out 100 MC simulations and in the remaining 27 MC simulations k_3 and k_2 were estimated together.

Table 13 shows the A-optimal settings that were selected. Note that, for all 100 MC simulations, the LO approach ranked K_{eq1} , k_2 , k_3 , k_4 , E_2 and ΔH and always left out k_6 , E_3 , E_5 and E_6 . The other five parameters were sometimes selected for consideration in **FIM_r**. For

example, k_5 was selected in 53 out of 100 MC simulations and E_{1f} was selected in 77 out 100 MC simulations. As shown in Table 13, the selected temperature settings for the new runs were at either lower or upper bound using the LO approach. However, for the PI approach, temperatures were sometimes at the bounds and sometimes and sometimes in between. Using the LO approach, the optimizer never chose to replicate the same experiment. However, using the PI approach, the optimizer selected the same settings for both new experiments 45% of time. Figure 8 compares SSD_θ values obtained by the LO and PI approaches for designing two sequential A-optimal experiments. As in scenario I, the LO approach provides superior results, on average compared to the PI approach.

Figure 8. Comparison of LO and PI approaches for designing two sequential A-optimal experiments for the scenario II

Table 13. A-optimal settings selected by LO and PI approaches and their frequencies

LO approach					
#	C _{SM10} (M)	C _{D0} (M)	C _{SM20} (M)	T (°C)	Frequency from 100 simulations
LO-A-1	1.350	0.080	1.500	45.000	50
	1.500	0.173	1.428	35.000	
LO-A -2	1.500	0.500	0.100	35.000	35
	0.818	0.134	1.500	35.000	
LO-A -3	1.500	0.500	0.100	45.000	15
	1.500	0.308	0.914	35.000	
PI approach					
#	C _{SM10} (M)	C _{D0} (M)	C _{SM20} (M)	T (°C)	Frequency from 100 simulations
PI-A-1	0.734	0.500	1.500	45.000	45
	0.734	0.500	1.500	45.000	
PI-A-2	0.815	0.069	1.459	44.770	26
	0.595	0.470	1.441	44.884	
PI-A-3	1.500	0.500	1.454	45.000	13
	0.724	0.500	1.500	40.000	
PI-A-4	1.318	0.244	1.341	37.210	6
	0.756	0.435	1.439	40.606	
PI-A-5	0.729	0.194	1.482	35.676	6
	1.486	0.294	1.497	35.000	
PI-A-6	0.760	0.406	1.500	45.00	2
	0.834	0.207	1.500	35.00	
PI-A-7	0.100	0.500	0.103	45.000	2
	0.778	0.500	1.500	45.000	

The LO and PI approaches were also used to design two new sequential V-optimal experiments (see Table 14). The LO approach resulted in three different sets of design settings with most popular conditions selected 52% of the time. No replicate experiments were selected. In all 100 MC simulations, the LO approach selected C_{D0} close to the lower bound and C_{SM20} at its upper bound. The selected temperatures were always at the lower or upper bound. The PI approach also resulted in three different sets of run conditions using the 100 MC simulations.

The second most popular settings (selected 32% of the time) resulted in a replicated experiment. Figure 9 compares boxplots for 100 values of SSD_{YW} obtained using the LO and PI approaches, indicating that the LO approach is superior to PI for this case study.

Table 14. V-optimal settings selected by LO and PI approaches and their frequencies

LO approach					
#	C _{SM10} (M)	C _{D0} (M)	C _{SM20} (M)	T (°C)	Frequency from 100 simulations
LO-V-1	1.2421	0.0166	1.5000	45.000	52
	1.5000	0.0250	1.5000	35.000	
LO-V-2	1.4873	0.0674	1.5000	45.000	38
	0.7710	0.0500	1.5000	45.000	
LO-V-3	1.3129	0.0377	1.5000	45.000	10
	0.8637	0.0585	1.5000	45.000	
PI approach					
#	C _{SM10} (M)	C _{D0} (M)	C _{SM20} (M)	T (°C)	Frequency from 100 simulations
PI-V-1	0.842	0.090	1.292	42.568	59
	0.624	0.139	1.500	35.000	
PI-V-2	0.814	0.047	1.500	45.000	32
	0.814	0.047	1.500	45.000	
PI-V-3	1.432	0.040	1.500	45.00	9
	0.761	0.382	1.500	45.00

Figure 9. Comparison on LO and PI approaches for designing two sequential V-optimal experiments for the scenario II

5.3 Comparing results from A- and V-optimal designs

Finally, A- and V-optimal designs were compared to illustrate the effectiveness of the LO methodology for their intended uses. As shown in Figure 10, the A-optimal designs produced better parameter estimates, on average, than the corresponding V-optimal designs. By contrast, the V-optimal designs resulted in better model predictions, on average, at the conditions of interest. The LO A- and V-optimal designs were both superior to a traditional approach (MS) where the modeler selects new conditions (randomly) from the corners of the operating space. Figure 11 summarizes the parameters that were estimated using the proposed LO A- and V-optimal MBDOE methodologies. In all 100 A-optimal design calculations, parameters $K_{eq1}, k_2, k_3, k_4, E_2$ and ΔH were estimated. Figure 11b shows that fewer parameters tended to be estimated from the V-optimal experiments, with K_{eq1}, k_2 , and ΔH estimated most of the time. Parameters k_6, E_3, E_5 and E_6 were never estimated from the either A- and V-optimal

experiments, presumably due to their relatively small influence on the model predictions and their highly correlated effects with other parameters. On average 6.0 parameters were estimated following the A-optimal experiments whereas only 4.9 parameters were estimated from the V-optimal experiments. This result is not surprising because the objective of the A-optimal design is to ensure that all the parameters are close to their true values, regardless of their influence on the model predictions whereas V-optimal design focuses on obtaining parameter values that lead to reliable predictions at the settings of interest.

In summary, the results indicate that it is worth using an optimizer to perform MBDoE calculations when the FIM is noninvertible (due to the limited data and the experimental resources). We recommend using the LO approach, rather than a PI approach, because better results were obtained on average, for the scenarios that were considered. A benefit of the proposed sequential MBDoE strategy, compared to traditional DoE is that any number of experiments can be designed, based on the model and the data that are currently available. Parameters that are not estimable from the available data are held at their initial guesses based on prior knowledge or scientific judgment. Additional parameters are automatically included in the estimation as additional information becomes available. After performing each round of experimentation, additional sequential experiments can be designed, and these experiments can be simulated to assess which additional parameters are likely to become estimable based on the anticipated data. Modelers who are interested in estimating parameters that are excluded from the parameter subset, can explore the causes of non-estimability (e.g., using identifiability techniques (Balsa-Canto et al., 2010; Petersen et al., 2001; McLean and McAuley, 2012)) and may find that additional measurements (e.g., concentrations of intermediate species) are required to obtain unique estimates of the non-estimable parameters. Based on the results in this

case study and from those provided in our previous paper, the LO approach seems to have superior results on average compared to the PI approach.

It is not possible for us to *prove* the effectiveness of the proposed LO approach using limited case studies. To confirm the general validity of these results, further testing with additional models and data sets is warranted. Comparisons should also be made with sequential Bayesian design methods. Given that D-optimal is the most commonly used MBDoE technique (Franceschini and Macchietto, 2008; John and Draper, 1975; Pinto et al., 1990; Thompson et al., 2010; Walter and Pronzato, 1990), it would be beneficial to evaluate the performance of the Leave Out approach for designing D-optimal experiments. Unfortunately, it is not straightforward to select a fair criterion to compare D-optimal LO and PI approaches. Recall that the SSD_{θ} criterion is a good match for A-optimality and the SSD_{Yw} criterion is a good match for V-optimality. When the **FIM** is noninvertible, the determinant of the true **FIM** is zero and there is no obvious way to compute a simple criterion related to the joint confidence region for the parameters. As a result, D-optimal designs were not considered in the current case study.

Figure 10. Boxplots for comparing (a) SSD_{θ} and (b) SSD_{γ_W} obtained by A- and V-optimal designs and MS approach.

Figure 11. Estimated parameters after using LO methodology to design two new experiments, based on 100 MC simulations.

6 CONCLUSIONS

In this article, two different methodologies are compared for dealing with a noninvertible Fisher Information Matrix (**FIM**) when designing A- and V-optimal experiments. The first methodology (the leave-out approach) uses a parameter subset selection technique to rank parameters and find problematic parameters that make the **FIM** noninvertible. By leaving out the problematic parameters, a reduced **FIM** can be used to design A- and V-optimal experiments.

The second methodology uses a pseudoinverse of the **FIM** in place of the noninvertible **FIM** during sequential A- and V- optimal design (PI method). Monte Carlo simulations were used to generate simulated data for each approach so that the LO and PI approaches could be compared using a nonlinear Michaelis Menten dynamic reaction model. Two scenarios were considered: I) operations at 40°C, II) operation at temperatures between 35 °C and 45 °C .

The LO and PI approaches were compared for both scenarios, revealing that the LO approach is superior on average for both A- and V-optimal sequential designs. The results confirm that, even when the FIM is non-invertible, A-optimal settings are more effective for improving the parameter estimates and V-optimal settings are more effective for improving model predictions at operating conditions of interest. We believe that the LO procedure described in this study will be helpful for designing experiments for chemical and biochemical processes with large number of parameters. Specifically, it will be helpful for performing automated experimental design calculations where numerical difficulties may arise due to noninvertibility of the **FIM**.

7 ACKNOWLEDGMENT

Authors acknowledge financial support from Eli Lilly and Company and technical advice from Dr. Salvador Garcia-Munoz and Dr. Brandon Jacob Reizman. An Ontario Trillium Scholarship for A. Shahmohammadi from government from the province of Ontario is also gratefully acknowledged.

8 NOMENCLATURE

Abbreviations

FIM	Fisher Information Matrix
FIM_r	Reduced FIM
FIM^{*-1}	Pseudoinverse of FIM
LO	Leave Out approach
PI	Pseudoinverse approach
MS	Modeler Selected approach
MBDoE	Model-Based Design of Experiment
MC	Monte Carlo Simulation
MSE	Mean-Squared Errors
ODE	Ordinary Differential Equation
SSD_θ	Sum of squared deviations between the parameter estimates their true values for parameter θ
SSD_{YW}	Sum of squared deviations between the model predictions at the conditions of interest and their noise free values

Symbols

C_{SM10}	Initial concentration of component SM1
C_{D0}	Initial concentration of component D
C_{SM20}	Initial concentration of component SM2
C_{SM1}	Time dependent concentration of component SM1
C_D	Time dependent concentration of component D
C_{SM2}	Time dependent concentration of component SM2
C_P	Time dependent concentration of component P
C_{H2O}	Time dependent concentration of component H2O
C_{SM1D}	Time dependent concentration of intermediate SM1D
C_{I1}	Time dependent concentration of impurity I1
C_{I2}	Time dependent concentration of impurity I2
C_{I3}	Time dependent concentration of impurity I3
C_{I4}	Time dependent concentration of impurity I4
C_i	Concentration of i th component in the reactor
d	Vector of settings for the decision variables
$\det(\cdot)$	Determinant
$\mathbf{g}(\cdot)$	General model prediction
g_i	Model prediction corresponds to the i th measurement
i	Integration counter for the measured responses
J	Objective function for parameter estimation
J_p	Objective function when all parameters are estimated
J_k	Objective function when k parameters are estimated
J_A	A-optimality criterion
J_D	D-optimality criterion
J_E	E-optimality criterion

J_V	V-optimality criterion
k & j	Iteration counter for the model parameters
k_j	j th reaction rate coefficient
K_{eq}	Equilibrium constant
L	Number of designed experiments
n	Number of run conditions
p	Number of parameters
r	Number sampling times
$\mathbf{R}_r$	Residual matrix
r_{cc}	Critical ratio criterion
$\mathbf{S}$	Sensitivity matrix
S_{ij}	Element of i th row and j th in sensitivity matrix
s_{θ_j}	Scaling factor for the initial uncertainties in the j th model parameters
s_{y_i}	Standard deviation of the of the j th measurement
$trace()$	Sum of diagonal elements of a matrix
v	Number different response variables
V_r	Volume of solution in semi-batch reactor
$\mathbf{W}$	Specific operating condition of interest
$\mathbf{X}$	Designed matrix for linear model
$\mathbf{X}_i$	Input settings corresponding to i th explanatory variables
x_j	j th correspond experimental setting for the explanatory variables
$\mathbf{X}_{old}$ or $\mathbf{Z}_{old}$	Initial or old designed matrix
$\mathbf{X}_{new}$ or $\mathbf{Z}_{new}$	New designed matrix
$\mathbf{X}_r$	Matrix of most-estimable parameters
$\mathbf{Y}$	Vector of N measured responses
y_j	j th measured response
$\mathbf{Z}$	Scaled sensitivity matrix
$\hat{\mathbf{Z}}_r$	Matrix of predictions for $\mathbf{Z}$
$(\cdot)^{-1}$	Inverse of a matrix

9 REFERENCES

- Asprey, S.P., Naka, Y., 1999. Mathematical problems in fitting kinetic models-some new perspectives. J. Chem. Eng. Japan. <https://doi.org/10.1252/jcej.32.328>
- Atkinson, A.C., 1982. Developments in the Design of Experiments, Correspondent Paper. Int. Stat. Rev. / Rev. Int. Stat. 50, 161–177. <https://doi.org/10.2307/1402599>
- Atkinson, A.C., Fedorov, V. V., 1975. The design of experiments for discriminating between two rival models. Biometrika. <https://doi.org/10.1093/biomet/62.1.57>

- Azzaro-Pantel, C., 2018. New Product Development and Supply Chains in the Pharmaceutical Industry. *Comput. Aided Chem. Eng.* 41, 1–26. <https://doi.org/10.1016/B978-0-444-63963-9.00001-4>
- Balsa-Canto, E., Alonso, A.A., Banga, J.R., 2010. An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Syst. Biol.* <https://doi.org/10.1186/1752-0509-4-11>
- Banga, J.R., Balsa-Canto, E., 2008. Parameter estimation and optimal experimental design. *Essays Biochem.* <https://doi.org/10.1042/bse0450195>
- Barz, T., López Cárdenas, D.C., Arellano-Garcia, H., Wozny, G., 2013. Experimental evaluation of an approach to online redesign of experiments for parameter determination. *AIChE J.* <https://doi.org/10.1002/aic.13957>
- Bauer, I., Bock, H.G., Körkel, S., Schlöder, J.P., 2000. Numerical methods for optimum experimental design in DAE systems. *J. Comput. Appl. Math.* [https://doi.org/10.1016/S0377-0427\(00\)00300-9](https://doi.org/10.1016/S0377-0427(00)00300-9)
- Ben-Zvi, A., McAuley, K., McLellan, J., 2004. Identifiability study of a liquid-liquid phase-transfer catalyzed reaction system. *AIChE J.* <https://doi.org/10.1002/aic.10202>
- Ben-Zvi, A., McLellan, P.J., McAuley, K.B., 2010. Identifiability of Non-Linear Differential Algebraic Systems via a Linearization Approach. *Can. J. Chem. Eng.* <https://doi.org/10.1002/cjce.5450840511>
- Benson, N., 2018. Quantitative Systems Pharmacology and Empirical Models: Friends or Foes? *CPT pharmacometrics Syst. Pharmacol.* <https://doi.org/10.1002/psp4.12375>
- Benyahia, B., Lakerveld, R., Barton, P.I., 2012. A plant-wide dynamic model of a continuous pharmaceutical process. *Ind. Eng. Chem. Res.* <https://doi.org/10.1021/ie3006319>
- Birtwistle, M.R., Mager, D.E., Gallo, J.M., 2013. Mechanistic Vs. empirical network models of drug action. *CPT Pharmacometrics Syst. Pharmacol.* <https://doi.org/10.1038/psp.2013.51>
- Boukouvala, F., Muzzio, F.J., Ierapetritou, M.G., 2011. Dynamic data-driven modeling of pharmaceutical processes. *Ind. Eng. Chem. Res.* <https://doi.org/10.1021/ie102305a>
- Box, G.E.P., Hunter, J.S., Hunter, W.G., 2005. *Statistics for Experimenters: design, innovation, and discovery.* John Wiley Sons, Inc. <https://doi.org/10.1017/CBO9781107415324.004>
- Box, G.E.P., Lucas, H.L., 2006. Design of Experiments in Non-Linear Situations. *Biometrika.* <https://doi.org/10.2307/2332810>
- Burth, M., 1999. Subset selection for improved parameter estimation in on-line identification of a synchronous Generator. *IEEE Trans. Power Syst.* 14, 218–225.
- Butcher, E.C., 2005. Can cell systems biology rescue drug discovery? *Nat. Rev. Drug Discov.*

<https://doi.org/10.1038/nrd1754>

- Casas-Orozco, D., Villa, A.L., Guerra, O.J., Reklaitis, G. V., 2018. Dynamic parameter estimation and identifiability analysis for heterogeneously-catalyzed reactions: Catalytic synthesis of nopol. *Chem. Eng. Res. Des.* <https://doi.org/10.1016/j.cherd.2018.04.002>
- Chaloner, K., Verdinelli, I., 1995. Bayesian Experimental Design: A Review. *Stat. Sci.* 10, 273–304. <https://doi.org/10.1214/ss/1177013604>
- Cho, K.H., Shin, S.Y., Kolch, W., Wolkenhauer, O., 2003. Experimental Design in Systems Biology, Based on Parameter Sensitivity Analysis Using a Monte Carlo Method: A Case Study for the TNF -Mediated NF- B Signal Transduction Pathway. *Simulation*. <https://doi.org/10.1177/0037549703040943>
- Chu, Y., Hahn, J., 2009. Parameter set selection via clustering of parameters into pairwise indistinguishable groups of parameters. *Ind. Eng. Chem. Res.* <https://doi.org/10.1021/ie800432s>
- Chu, Y., Huang, Z., Hahn, J., 2009. Improving prediction capabilities of complex dynamic models via parameter selection and estimation. *Chem. Eng. Sci.* 64, 4178–4185. <https://doi.org/10.1016/j.ces.2009.06.057>
- Cui, W.J., McAuley, K.B., Spence, R.E., Xie, T., 2015. Mathematical Modeling of Polyether Production from 1,3-Propanediol: Accounting for Linear Oligomers. *Macromol. React. Eng.* <https://doi.org/10.1002/mren.201400038>
- Cui, W.J., McAuley, K.B., Whitney, R.A., Spence, R.E., Xie, T., 2013. Mathematical model of polyether production from 1,3-Propanediol. *Macromol. React. Eng.* 7, 237–253. <https://doi.org/10.1002/mren.201200068>
- Detle, H., Wong, W.K., 1999. E-optimal designs for the Michaelis–Menten model. *Stat. Probab. Lett.* [https://doi.org/10.1016/S0167-7152\(99\)00033-4](https://doi.org/10.1016/S0167-7152(99)00033-4)
- DiMasi, J.A., Hansen, R.W., Grabowski, H.G., 2003. The price of innovation: New estimates of drug development costs. *J. Health Econ.* [https://doi.org/10.1016/S0167-6296\(02\)00126-1](https://doi.org/10.1016/S0167-6296(02)00126-1)
- Domagalski, N.R., Mack, B.C., Tabora, J.E., 2015. Analysis of Design of Experiments with Dynamic Responses. *Org. Process Res. Dev.* <https://doi.org/10.1021/acs.oprd.5b00143>
- Durán, M.A., White, B.S., 1995. Bayesian estimation applied to effective heat transfer coefficients in a packed bed. *Chem. Eng. Sci.* 50, 495–510.
- Eghtesadi, Z., McAuley, K.B., 2016. Mean-squared-error-based method for parameter ranking and selection with noninvertible fisher information matrix. *AIChE J.* <https://doi.org/10.1002/aic.15096>
- Eghtesadi, Z., McAuley, K.B., 2014. Mean square error based method for parameter ranking and selection to obtain accurate predictions at specified operating conditions. *Ind. Eng. Chem.*

- Res. <https://doi.org/10.1021/ie5002444>
- Eghtesadi, Z., Wu, S., McAuley, K.B., 2013. Development of a model selection criterion for accurate model predictions at desired operating conditions. *Ind. Eng. Chem. Res.* <https://doi.org/10.1021/ie302408b>
- Eldén, L., 1982. A weighted pseudoinverse, generalized singular values, and constrained least squares problems. *BIT*. <https://doi.org/10.1007/BF01934412>
- Espie, D.M., Macchietto, S., 1988. Nonlinear Transformations for Parameter Estimation. *Ind. Eng. Chem. Res.* <https://doi.org/10.1021/ie00083a037>
- FDA, 2004. Pharmaceutical CGMPs for the 21s Century - A risk-based approach. FDA.
- Ford, I., Silvey, S.D., 1980. A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika* 67, 381–388.
- Ford, I., Titterton, D.M., Kitsos, C.P., 1989. Recent advances in nonlinear experimental design. *Technometrics*. <https://doi.org/10.1080/00401706.1989.10488475>
- Franceschini, G., Macchietto, S., 2008. Model-based design of experiments for parameter precision: State of the art. *Chem. Eng. Sci.* <https://doi.org/10.1016/j.ces.2007.11.034>
- François, N., Govaerts, B., Boulanger, B., 2004. Optimal designs for inverse prediction in univariate nonlinear calibration models, in: *Chemometrics and Intelligent Laboratory Systems*. pp. 283–292. <https://doi.org/10.1016/j.chemolab.2004.05.005>
- Gernaey, K. V., Gani, R., 2010. A model-based systems approach to pharmaceutical product-process design and analysis. *Chem. Eng. Sci.* <https://doi.org/10.1016/j.ces.2010.05.003>
- Goos, P., Syafitri, U., 2014. V-optimal mixture designs for the qth degree model. *Chemom. Intell. Lab. Syst.* 136, 173–178. <https://doi.org/10.1016/j.chemolab.2014.04.019>
- Greville, T.N.E., 1959. The Pseudoinverse of a Rectangular or Singular Matrix and Its Application to the Solution of Systems of Linear Equations. *SIAM Rev.*
- Hallow, D.M., Mudryk, B.M., Braem, A.D., Tabora, J.E., Lyngberg, O.K., Bergum, J.S., Rossano, L.T., Tummala, S., 2010. An example of utilizing mechanistic and empirical modeling in quality by design. *J. Pharm. Innov.* <https://doi.org/10.1007/s12247-010-9094-y>
- Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 55–67.
- Holmberg, A., 1982. On the practical identifiability of microbial growth models incorporating Michaelis-Menten type nonlinearities. *Math. Biosci.* [https://doi.org/10.1016/0025-5564\(82\)90061-X](https://doi.org/10.1016/0025-5564(82)90061-X)
- Huang, J., Kaul, G., Cai, C., Chatlapalli, R., Hernandez-Abad, P., Ghosh, K., Nagi, A., 2009.

- Quality by design case study: An integrated multivariate approach to drug product and process development. *Int. J. Pharm.* <https://doi.org/10.1016/j.ijpharm.2009.07.031>
- Issanchou, S., Cognet, P., Cabassud, M., 2005. Sequential experimental design strategy for rapid kinetic modeling of chemical synthesis. *AIChE J.* 51, 1773–1781.
- Issanchou, S., Cognet, P., Cabassud, M., 2003. Precise parameter estimation for chemical batch reactions in heterogeneous medium. *Chem. Eng. Sci.* 58, 1805–1813.
- Jain, S., 2014. Quality by design (QBD): A comprehensive understanding of implementation and challenges in pharmaceuticals development. *Int. J. Pharm. Pharm. Sci.* <https://doi.org/undefined>
- Jayasankar, B.R., Ben-Zvi, A., Huang, B., 2009. Identifiability and estimability study for a dynamic solid oxide fuel cell model. *Comput. Chem. Eng.* 33, 484–492.
- Johansen, T.A., 1997. On {T}ikhonov Regularization, Bias and Variance in Nonlinear System Identification. *Automatica* 33, 441–446.
- John, R.C.S., Draper, N.R., 1975. D-Optimality for regression designs: A review. *Technometrics* 17, 15–23. <https://doi.org/10.1080/00401706.1975.10489266>
- Karimi, H., Cowperthwaite, E. V., Olayiwola, B., Farag, H., McAuley, K.B., 2017. Modelling of heat transfer and pyrolysis reactions in an industrial ethylene cracking furnace. *Can. J. Chem. Eng.* <https://doi.org/10.1002/cjce.22844>
- Karimi, H., Schaffer, M.A., McAuley, K.B., 2012. A Kinetic Model for Non-Oxidative Thermal Degradation of Nylon 66. *Macromol. React. Eng.* 6, 93–109.
- Kou, B., McAuley, K.B., Hsu, J.C.C., Bacon, D.W., 2005a. Mathematical model and parameter estimation for gas-phase ethylene/hexene copolymerization with metallocene catalyst. *Macromol. Mater. Eng.* 290, 537–557. <https://doi.org/10.1002/mame.200400392>
- Kou, B., McAuley, K.B., Hsu, J.C.C., Bacon, D.W., 2005b. Mathematical model and parameter estimation for gas-phase ethylene/hexene copolymerization with metallocene catalyst. *Macromol. Mater. Eng.* <https://doi.org/10.1002/mame.200400392>
- Kravaris, C., Hahn, J., Chu, Y., 2013. Advances and selected recent developments in state and parameter estimation. *Comput. Chem. Eng.* 51, 111–123.
- Lehninger, 2009. Principles of Biochemistry, 6th ed., Journal of Chemical Information and Modeling. <https://doi.org/10.1017/CBO9781107415324.004>
- Li, Y.H., Yeh, P.C., 2012. An interpretation of the moore-penrose generalized inverse of a singular fisher information matrix. *IEEE Trans. Signal Process.* 60, 5532–5536. <https://doi.org/10.1109/TSP.2012.2208105>
- Littlejohns, J. V, McAuley, K.B., Daugulis, A.J., 2010. Model for a solid–liquid stirred tank two-

- phase partitioning bioscrubber for the treatment of BTEX. *J. Hazard. Mater.* 175, 872–882.
- Liu, S., Neudecker, H., 1995. A V-optimal design for Scheffé's polynomial model. *Stat. Probab. Lett.* 23, 253–258. [https://doi.org/10.1016/0167-7152\(94\)00122-O](https://doi.org/10.1016/0167-7152(94)00122-O)
- Ljung, L., Glad, T., 1994. On global identifiability for arbitrary model parametrizations. *Automatica*. [https://doi.org/10.1016/0005-1098\(94\)90029-9](https://doi.org/10.1016/0005-1098(94)90029-9)
- López C., D.C., Barz, T., Körkel, S., Wozny, G., 2015. Nonlinear ill-posed problem analysis in model-based parameter estimation and experimental design. *Comput. Chem. Eng.* 77, 24–42.
- López C, D.C., Barz, T., Peñuela, M., Villegas, A., Ochoa, S., Wozny, G., 2013. Model-based identifiable parameter determination applied to a simultaneous saccharification and fermentation process model for bio-ethanol production. *Biotechnol. Prog.* 29, 1064–82.
- Lund, B.F., Foss, B.A., 2008. Parameter ranking by orthogonalization-Applied to nonlinear mechanistic models. *Automatica* 44, 278–281. <https://doi.org/10.1016/j.automatica.2007.04.006>
- Mallows, C.L., 1973. Some Comments on Cp. *Technometrics* 15, 661. <https://doi.org/10.2307/1267380>
- Maria, G., 2004. A review of algorithms and trends in kinetic model identification for chemical and biochemical systems. *Chem. Biochem. Eng. Q.* 18, 195–222.
- McLean, K.A.P., McAuley, K.B., 2012. Mathematical modelling of chemical processes-obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *Can. J. Chem. Eng.* 90, 351–366.
- Meshram, M., Naderi, S., McConkey, B., Ingalls, B., Scharer, J., Budman, H., 2013. Modeling the coupled extracellular and intracellular environments in mammalian cell culture. *Metab. Eng.* 19, 57–68.
- Murphy, E.F., Gilmour, S.G., Crabbe, M.J.C., 2004. Efficient and cost-effective experimental determination of kinetic constants and data: The success of a Bayesian systematic approach to drug transport, receptor binding, continuous culture and cell transport kinetics. *FEBS Lett.* 556, 193–198.
- Nayak, R.P., Foudriat, E.C., 1974. Sequential parameter estimation using pseudoinverse. *IEEE Trans. Automat. Contr.* AC19, 81–83.
- Neshat, S.S., Pope, G.A., 2017. Compositional Three-Phase Relative Permeability and Capillary Pressure Models Using Gibbs Free Energy, in: *SPE Reservoir Simulation Conference*. <https://doi.org/10.2118/182592-MS>
- Ngo, V. V., Latifi, M.A., Simonnot, M.O., 2013. Estimability Analysis and Optimisation of Soil Hydraulic Parameters from Field Lysimeter Data. *Transp. Porous Media* 98, 485–504.

- Nguyen, T.T.H., Teratani, S., Tanaka, R., Endo, A., Hirao, M., 2017. A framework for developing a structure-based lumping kinetic model for the design and simulation of refinery reactors. *Comput. Chem. Eng.* 106, 385–395.
- Pandey, P., Bharadwaj, R., 2016. *Predictive Modeling of Pharmaceutical Unit Operations*. Woodhead Publishing.
- Petersen, B., Gernaey, K., Vanrolleghem, P.A., 2001. Practical identifiability of model parameters by combined respirometric-titrimetric measurements, in: *Water Science and Technology*.
- Petrides, D., Koulouris, A., Siletti, C., Jiménez, J.O., Lagonikos, P.T., 2010. The Role of Simulation and Scheduling Tools in the Development and Manufacturing of Active Pharmaceutical Ingredients, in: *Chemical Engineering in the Pharmaceutical Industry: R&D to Manufacturing*. <https://doi.org/10.1002/9780470882221.ch28>
- Petrides, D.P., Koulouris, A., Lagonikos, P.T., 2002. The role of process simulation in pharmaceutical process development and product commercialization. *Pharm. Eng.* 22, 56–65.
- Pinto, J.C., Lobão, M.W., Monteiro, J.L., 1990. Sequential experimental design for parameter estimation: a different approach. *Chem. Eng. Sci.* 45, 883–892.
- Pohjanpalo, H., 1978. System identifiability based on the power series expansion of the solution. *Math. Biosci.* [https://doi.org/10.1016/0025-5564\(78\)90063-9](https://doi.org/10.1016/0025-5564(78)90063-9)
- Poós, T., Szabó, V., 2017. Application of Mathematical Models Using Volumetric Transfer Coefficients in Fluidized Bed Dryers, in: *Energy Procedia*. <https://doi.org/10.1016/j.egypro.2017.03.1079>
- Quaglio, M., Fraga, E.S., Galvanin, F., 2018. Model-based design of experiments in the presence of structural model uncertainty: an extended information matrix approach. *Chem. Eng. Res. Des.* <https://doi.org/10.1016/j.cherd.2018.04.041>
- Quaglio, M., Waldron, C., Pankajakshan, A., Cao, E., Gavrilidis, A., Fraga, E.S., Galvanin, F., 2019. An online reparametrisation approach for robust parameter estimation in automated model identification platforms. *Comput. Chem. Eng.* <https://doi.org/10.1016/j.compchemeng.2019.01.010>
- Rao, C.R., 1966. Linear statistical inference and its applications, *USSR Computational Mathematics and Mathematical Physics*. [https://doi.org/10.1016/0041-5553\(66\)90024-3](https://doi.org/10.1016/0041-5553(66)90024-3)
- Ruggoo, A., Vandebroek, M., 2004. Bayesian sequential script design optimal model-robust designs. *Comput. Stat. Data Anal.* 47, 655–673. <https://doi.org/10.1016/j.csda.2003.09.014>
- Shahmohammadi, A., McAuley, K.B., 2019. Sequential Model-Based A-Optimal Design of Experiments When the Fisher Information Matrix Is Noninvertible. *Ind. Eng. Chem. Res.* 58, 1244–1261. <https://doi.org/10.1021/acs.iecr.8b03047>

- Shirt, R.W., Harris, T.J., Bacon, D.W., 1994. Experimental Design Considerations for Dynamic Systems. *Ind. Eng. Chem. Res.* 33, 2656–2667.
- Srinath, S., Gunawan, R., 2010. Parameter identifiability of power-law biochemical system models. *J. Biotechnol.* <https://doi.org/10.1016/j.jbiotec.2010.02.019>
- Stigler, S.M., 1971. Optimal experimental design for polynomial regression. *J. Am. Stat. Assoc.* 66, 311–318.
- Thompson, D.E., McAuley, K.B., McLellan, P.J., 2010. Design of optimal sequential experiments to improve model predictions from a polyethylene molecular weight distribution model. *Macromol. React. Eng.* 4, 73–85.
- Thompson, D.E., McAuley, K.B., McLellan, P.J., 2009. Parameter estimation in a simplified MWD model for HDPE produced by a ziegler-natta catalyst. *Macromol. React. Eng.* 3, 160–177.
- Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* 73, 273–282.
- Tikhonov, A.N., Goncharsky, A. V., Stepanov, V. V., Yagola, A.G., 1978. Numerical methods for the solution of ill-posed problems. *Math. Comput.* 32, 1320. <https://doi.org/10.2307/2006360>
- Vajda, Sandor, Godfrey, K.R., Rabitz, H., 1989. Similarity transformation approach to identifiability analysis of nonlinear compartmental models. *Math. Biosci.* [https://doi.org/10.1016/0025-5564\(89\)90024-2](https://doi.org/10.1016/0025-5564(89)90024-2)
- Vajda, S., Rabitz, H., Walter, E., Lecourtier, Y., 1989. Qualitative and Quantitative Identifiability Analysis of Nonlinear Chemical Kinetic Models. *Chem. Eng. Commun.* 83, 191–219.
- Velez-Reyes, M., Verghese, G.C., 1995. Subset selection in identification, and application to speed and parameter estimation for induction machines. *Control Appl.* 991–997.
- Walter, E., Lecourtier, Y., 1982. Global approaches to identifiability testing for linear and nonlinear state space models. *Math. Comput. Simul.* [https://doi.org/10.1016/0378-4754\(82\)90645-0](https://doi.org/10.1016/0378-4754(82)90645-0)
- Walter, E., Pronzato, L., 1990. Qualitative and quantitative experiment design for phenomenological models-A survey. *Automatica* 26, 195–213.
- Welch, W.J., 1984. Computer-aided design of experiments for response estimation. *Technometrics* 26, 217–224.
- Woloszyn, J.D., Hesse, P., Hungenberg, K.D., McAuley, K.B., 2013. Parameter selection and estimation techniques in a styrene polymerization model. *Macromol. React. Eng.* <https://doi.org/10.1002/mren.201200074>

- Wong, W.K., 1994. Comparing robust properties of A, D, E and G-optimal designs. *Comput. Stat. Data Anal.* [https://doi.org/10.1016/0167-9473\(94\)90161-9](https://doi.org/10.1016/0167-9473(94)90161-9)
- Wu, S., Harris, T.J., McAuley, K.B., 2008. The Use of Simplified or Misspecified Models: Linear Case. *Can. J. Chem. Eng.* 85, 386–398.
- Wu, S., McAuley, K.B., Harris, T.J., 2011. Selection of simplified models: II. Development of a model selection criterion based on mean squared error. *Can. J. Chem. Eng.* 89, 325–336.
- Yao, K.Z., Shaw, B.M., Kou, B., McAuley, K.B., Bacon, D.W., 2003. Modeling Ethylene/Butene Copolymerization with Multi- site Catalysts: Parameter Estimability and Experimental Design. *Polym. React. Eng.* 11, 563–588.
- Yu, L.X., Amidon, G., Khan, M.A., Hoag, S.W., Polli, J., Raju, G.K., Woodcock, J., 2014. Understanding Pharmaceutical Quality by Design. *AAPS J.* <https://doi.org/10.1208/s12248-014-9598-3>
- Yue, H., Brown, M., Knowles, J., Wang, H., Broomhead, D.S., Kell, D.B., 2006. Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: A case study of an NF- κ B signalling pathway. *Mol. Biosyst.* <https://doi.org/10.1039/b609442b>
- Zhang, Y., Edgar, T.F., 2008. PCA combined model-based design of experiments (DOE) criteria for differential and algebraic system parameter estimation. *Ind. Eng. Chem. Res.* 47, 7772–7783. <https://doi.org/10.1021/ie071206c>
- Zhao, Y.R., Arriola, D.J., Puskas, J.E., McAuley, K.B., 2017. Applying multidimensional method of moments for modeling and estimating parameters for arborescent polyisobutylene production in batch reactor. *Macromol. Theory Simulations* 26.
- Zhao, Y.R., McAuley, K.B., Puskas, J.E., Dos Santos, L.M., Alvarez, A., 2013. Mathematical modeling of arborescent polyisobutylene production in batch reactors. *Macromol. Theory Simulations* 22, 155–173. <https://doi.org/10.1002/mats.201200058>