DoE-integrated Sparse Identification of Nonlinear Dynamics for Automated Model Generation and Parameter Estimation in Kinetic Studies

Wenyao Lyu*a*, Federico Galvanin*a*

aDepartment of Chemical Engineering, University College London (UCL), Torrington Place, London, WC1E 7JE, United Kingdom

f.galvanin@ucl.ac.uk

Abstract

Digital twins have revolutionized manufacturing by utilising robust kinetic models to predict the behaviour of biochemical reaction systems under variable operating conditions. Identifying accurate expressions for reaction system models formulated as sets of differential and algebraic equations (DAEs) is challenging due to numerous state variables and kinetic parameters, compounded by limited observations and experimental errors. To enhance reliability, model structure confirmation should precede parameter estimation and validation. Traditional model-building approaches require prior knowledge of candidate models, whereas model generation methods like Sparse Identification of Nonlinear Dynamics (SINDy) only require the definition of a library of function terms. In this paper, we propose a new model identification framework, named Design of Experiment-integrated Sparse Identification of Nonlinear Dynamics (DoE-SINDy) to iteratively generate, evaluate, and select candidate models to represent systems with minimal training data. Tested on a simulated case study, DoE-SINDy succeeds in identifying an assumed true model efficiently with a limited experimental budget and noisy datasets.

**Keywords**: model identification, model generation, kinetic studies, sparse regression, data-driven modelling

* 1. Introduction

Digitisation is driving a profound transformation in manufacturing sectors by implementing digital twins based on robust kinetic models for representing the dynamic behaviour of reaction systems modelled as a set of differential and algebraic equations. Identifying these kinetic models from limited observations is very challenging due to their complex model structure, numerous variables, and parameters (Quaglio et al., 2020). Constructing a suitable model for digital twins often requires costly experimentation and a substantial investment of time and analytical resources. Furthermore, the limited observations and inevitable experimental errors further hinder the adequate identification.

The reliability of identified models is determined by the adequacy of model structure (i.e., set of equations) and parameter precision. Model structure adequacy is a statistical measure of how well a model fits the data, while parameter precision is related to the uncertainty of parameter estimates. Thus, it is necessary to identify a model with an accurate model structure and reduce the uncertainty in the parameters as much as possible.

Traditional model-building approaches involve model selection for identifying the most accurate model structure from a set of known model structures (Asprey and Macchietto, 2000), and model modification for correcting the unsuitable structure in the selected model. Challenges may arise when there is insufficient theoretical understanding of the system to generate candidate models or when the 'true' (i.e., most suitable) model is not among the candidate models, a common scenario in chemical and biochemical processes. Model generation approaches such as SINDy (Brunton et al., 2016) and AI-DARWIN (Chakraborty et al., 2021) are affected by several limitations: 1) potential generation of unidentifiable models; 2) absence of a rigorous model evaluation stage; 3) model generation is extremely sensitive to experimental conditions explored in the training set; 4) lack of a model selection stage. Motivated by these challenges, a new systematic model identification framework named DoE-SINDy is proposed in this paper for simultaneously generating model structures and estimating parameters from a small dataset to represent a dynamic system. To realise it, this framework aims to integrate model generation techniques and design of experiments (DoE) to expedite the identification of kinetic models.

* 1. Methodology

A dynamic chemical reaction model can be formulated as a set of differential and algebraic equations (DAEs) in the form

|  |  |
| --- | --- |
| $f\left(\dot{x},x,u,t,θ\right)=0$  | (1) |
| $\hat{x}=h(u, t, θ)$ |

where $f$ and $h$ are respectively an $N\_{f} $and $N\_{x} $vector of model equations, $x$ is an $N\_{x} $vector of state variables, $u\in U$ is an $N\_{u}$ vector of control input variables, *t* is time and $θ$ is an $N\_{θ}$vector of model parameters, $\hat{x}$ represents an $N\_{x}$ vector of model predictions for a measurable set of system state variables $x$. This research focuses on simultaneously identifying the model structure $f$**,** and parameters$θ$.



Figure 1 Framework of DoE-SINDy for identifying the most suitable models from experimental data.

SINDy is a main sparse regression-based approach that is capable of deriving models with minimal prior knowledge of physical mechanisms. As demonstrated in the literature by Brunton et al. (2016), SINDy operates under the assumption that a process model can be efficiently represented by only a few important terms that govern the dynamics. It adeptly integrates the most suitable terms from a candidate term library $F\left(X\right)$ and estimates parameters $θ$ to formulate model expressions, which strikes an equilibrium between model accuracy and complexity. Thus, it averts over-fitting while preserving the underlying physics of the discovered model. DoE-SINDy begins with a Design of Experiments (DoE), where operating conditions are specified based on underlying physical constraints (block 1 in Figure 1). Subsequently, time-varying measurements, such as concentrations of specific components, are collected along the trajectory, and the time derivatives of these measurements are numerically approximated due to their continuity in time (blocks 2 and 3 in Figure 1). As an extension of the original SINDy, DoE-SINDy utilises the original SINDy for model generation (blocks 4, 5 in Figure 1), employing an open-source library called *PySINDy* (de Silva et al., 2020; Kaptanoglu et al., 2022). In contrast to the original SINDy, DoE-SINDy is a closed-loop framework, iteratively generating multiple models from increasing-size datasets to ensure the identification of an assumed true model as shown in Figure 1. Additionally, DoE-SINDy incorporates a preliminary analysis stage (blocks 6, 7, 8 in Figure 1) to reject unidentifiable models and those with excessive terms. Identifiability, as defined in Miao et al., (2011), refers to the property of a dynamic system where the parameters $θ$ can be uniquely determined from the given system input $u$ and the measurable system output $x$. DoE-SINDy implements sensitivity-based practical identifiability analysis based on the calculation of Fisher Information Matrix (FIM), denoted as $H\_{θ}$. When analysing the identifiability over a trajectory, according to Waldron et al. (2019), $H\_{θ}$ is computed as the summation of $H\_{θ}\left(t\_{m}\right)$ at each sampling point

|  |  |
| --- | --- |
| $\left[H\_{θ}\right]\_{kl}≈\sum\_{m=1}^{N\_{m}}\frac{1}{σ\_{n}^{2}}q\_{nk}\left(t\_{m}^{i}\right)q\_{nl}\left(t\_{m}^{i}\right), for k,l=1,…,N\_{θ}$  | (2) |

In (2), $q\_{nk}\left(t\_{m}\right)$ is the *nk*-th element of the sensitivity $Q $representing the sensitivity of the response $x\_{n}$ to the parameter $θ\_{k}$ at time sampling point $t\_{m}$, calculated according to Franceschini et al. (2008). When $H\_{θ}$ is full rank, the model is considered locally identifiable. Conversely, when $H\_{θ}$ is singular, this indicates that the model is not identifiable (Miao et al., 2011). Since the assumption of noise distribution is not considered when implementing Sequentially Thresholded Least-squares algorithm (STLSQ) to estimate the parameters in the model generation section. DoE-SINDy incorporates a Maximum Likelihood (ML) estimator to robustly re-estimate model parameters (block 9 in Figure 1) under the assumption of measurements corrupted by Gaussian noise with zero mean and constant standard deviation $σ$.

Subsequentially, DoE-SINDy includes a model evaluation stage for selecting accurate model(s) from the generated models via a $χ^{2}$ test (block 10 in Figure 1) and assessing parameter precision via a *t*-test (block 11 in Figure 1).

* 1. Model description

We test the performance of DoE-SINDy to the same case study used in Quaglio et al. (2020). The goal is to identify an assumed true kinetic model of a batch reaction system described in equation (3).

|  |  |
| --- | --- |
| $\frac{dC\_{A}}{dt}=-k\_{1}C\_{A}^{2}$ $\frac{dC\_{B}}{dt}=k\_{1}C\_{A}^{2}-k\_{2}C\_{B}$ $\frac{dC\_{C}}{dt}=k\_{2}C\_{B}$  | (3) |

In equation (3), the reaction rate constants are denoted by $k\_{1}$ and $k\_{2}$. To collect experimental data in silico, we specified various initial concentrations of A within the range from 40 to 250 mol m3 s-1 to simulate the process multiple times. The duration of each experiment is 350 s, and measurements are taken from 100 to 300 s every 10 s. 20 experiments are simulated for data collection, and the initial concentration of species A is designed based on a Latin Hypercube Sampling (LHS) experimental design. Gaussian distributed noise with zero mean and constant standard deviation $σ=0.1$ mol m-3 is added to the simulated data to generate in-silico measurements.

* 1. Results

The performance of the model generation of DoE-SINDy is presented here along with a discussion of the findings and a comparison between the efficiency of identifying the true model between DoE-SINDy and the original SINDy.

*4.1 Algorithm efficiency using a specific LHS dataset*

The structure and parameters of the model generated through iterations are significantly influenced by the training data. Even when using identical measurements from the same 20 experiments, the generated models may exhibit diversity.

To illustrate how the order of data loading impacts the model structure in each iteration, we examine four different data loading sequences α, β, γ, and δ. It is worth noting that, SINDy estimates the parameters of every term in a model, as there are 4 terms in the true model, the number of parameters $N\_{θ}$ of the SINDy generated model having the identical model structure as the true model is 4 ($θ\_{1}$, $θ\_{2}, θ\_{3} and θ\_{4}$) instead of 2 ($k\_{1}$ and $k\_{2}$). To simply visualise the models generated over the iterative process, $N\_{θ}$ is used to represent the model structure.

As illustrated in Figure 2, in general, with the increase of number of iterations, $N\_{θ}$ decreases, eventually converging to 5 parameters. The true model with 4 parameters is successfully identified at one of the iterations during the model identification process, but only with sequences α and γ. These observations show that the training data used is crucial in determining whether the true model is identified.

Identifiability results show that the generated models are unidentifiable, as DoE-SINDy generated models have anti-correlated parameters, i.e., $θ\_{1}=k\_{1}$ and $θ\_{2}=-k\_{1}$, $θ\_{3}=-k\_{2}$ and $θ\_{4}=k\_{2}$. Because of this, the parameters cannot be precisely estimated.

The results from $χ^{2}$ test show that a model is falsified for under-fitting even when the true model structure is identified. For example, the $χ^{2}$ of the model with 4 parameters identified with sequence γ is $χ^{2}=1.81×10^{4} $which is larger than the reference $χ\_{0.95}^{2}=8.16×10^{2}$. However, there are negligible deviations between the measurements and the predictions. In that case, the true model may not be selected out of the generated models according to the results of $χ^{2}$ test.



Figure 2 Number of parameters of the generated model at each iteration based on 4 different experimental data loading sequences (a) α, (b) β, (c) γ, and (d) δ. The red line refers to the true number of parameters $N\_{θ}$, which is 4, bars refer to the defined initial concentration of A in each LHS experiment. The line illustrates $N\_{θ}$ for the model generated at each iteration.

*4.2 Capability in general conditions and comparison with SINDy*

To compare original SINDy and DoE-SINDY we investigate four different data sizes (5, 10, 15 and 20 experiments). For each data size, 20 datasets are generated. Within a single dataset, LHS is employed to design the initial concentrations of A for each experiment, and the code DoE-SINDy is executed 100 times with shuffled sequences to mitigate the impact of sequence on model generation. The Overall Success Rate, denoted as $η\_{OSR}$, and defined as the ratio between ‘successful datasets’ out of 20 datasets at this data size, is introduced as a metric to quantify the efficiency in identifying the true model structure.

Table 1 shows that the original SINDy only succeeds in identifying the true model within a dataset obtained from 10 experiments, while fails in all datasets obtained from 5, 15, and 20 experiments. Conversely, DoE-SINDy consistently succeeds with nearly all datasets at the data sizes of 10, 15 and 20 experiments. This observation reveals that by identifying models iteratively can significantly enhance the efficiency of identifying the true model. Additionally, it is beneficial to explore more sequences before conducting additional experiments.

Table 1 Comparison of the efficiency in identifying the true model between original SINDy and DoE-SINDy.

|  |  |
| --- | --- |
| Number of Experiments | $$η\_{OSR}/100\%$$ |
| SINDy | DoE-SINDy |
| 20 | 0 | 100 |
| 15 | 0 | 95 |
| 10 | 5 | 95 |
| 5 | 0 | 0 |

* 1. Conclusions

This research developed a novel model identification framework (DoE-SINDy) for automatically generating the most suitable kinetic model for dynamic processes in bio(chemical) systems. By combining iterative model generation via SINDy and model evaluation based on identifiability, goodness-of-fit and parameter precision, DoE-SINDy overcomes the challenges associated with limited and noisy observations and lack of the full expression of a model, significantly enhancing the model identification efficiency of the original SINDy. DoE-SINDy can be a promising tool for the identification of reliable and robust kinetic models in bio(chemical) reaction systems with a limited number of runs.

Acknowledgements

The authors gratefully acknowledge the funding support from UCL EPSRC DTP (EP/R513143/1 and EP/W524335/1).

References

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

S.L. Brunton, J.L. Proctor, and J.N. Kutz, 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the national academy of sciences, 113(15), pp.3932-3937.

A. Chakraborty, A. Sivaram, and V. Venkatasubramanian, 2021. AI-DARWIN: A first principles-based model discovery engine using machine learning. Computers & Chemical Engineering, 154, p.107470.

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

A.A. Kaptanoglu, B.M. de Silva, U. Fasel, K. Kaheman, A.J. Goldschmidt, J.L. Callaham, C.B. Delahunt, Z.G. Nicolaou, K. Champion, J.C. Loiseau, and J.N. Kutz, 2021. PySINDy: A comprehensive Python package for robust sparse system identification. arXiv preprint arXiv:2111.08481.

H. Miao, X. Xia, A.S. Perelson and H. Wu, 2011. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM review, 53(1), pp.3-39.

M. Quaglio, L. Roberts, M.S.B. Jaapar, E.S. Fraga, V. Dua, and F. Galvanin, 2020. An artificial neural network approach to recognise kinetic models from experimental data. Computers & Chemical Engineering, 135, p.106759.

C. Waldron, A. Pankajakshan, M. Quaglio, E. Cao, F. Galvanin, and A. Gavriilidis, 2019. Closed-loop model-based design of experiments for kinetic model discrimination and parameter estimation: benzoic acid esterification on a heterogeneous catalyst. Industrial & Engineering Chemistry Research, 58(49), pp.22165-22177.