Novel techniques for kinetic model identification and improvement

Marco Quaglio

A Thesis submitted in partial fulfilment

of the requirements for the

Doctor of Philosophy

of

University College London.

Department of Chemical Engineering

University College London

February 1, 2020

I, Marco Quaglio, confirm that the work presented in this Thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work. The realisation of this Thesis has involved the intellectual support of other people to whom I am extremely grateful. Most of the material presented in this work has been already published or submitted for undergoing a peer reviewing process in the following.

PUBLICATIONS IN INTERNATIONAL JOURNALS

Quaglio M., Fraga E. S., Galvanin F., Model-based design of experiments in the presence of structural model uncertainty: an extended information matrix approach, *Chemical Engineering Research and Design* 136, 2018, pp. 129-143

Quaglio M., Fraga E. S., Cao E., Gavriilidis A., Galvanin F., A model-based data mining approach for determining the domain of validity of approximated models, *Chemometrics and Intelligent Laboratory Systems* 172, 2018, pp. 58-67

Quaglio M., Waldron C., Pankajakshan A., Cao E., Gavriilidis, A., Fraga E. S., Galvanin F., On the use of online reparametrization in automated platforms for kinetic model identification, *Chemie Ingenieur Technik* 91(3), 2019, pp. 268-276

Quaglio M., Waldron C., Pankajakshan A., Cao E., Gavriilidis A., Fraga E. S., Galvanin F., An online reparametrisation approach for robust parameter estimation in automated model identification platforms, *Computers & Chemical Engineering* 124, 2019, pp. 270-284

Quaglio M., Bezzo F., Gavriilidis A., Cao E., Al-Rifai N., Galvanin F., Identification of kinetic models of methanol oxidation on silver in the presence of uncertain catalyst behaviour, *AIChE Journal* 65(10), 2019, pp:e16707

Quaglio M., Fraga E. S., Galvanin F., A diagnostic procedure for improving the structure of approximated kinetic models, *Computers & Chemical Engineering*, 2019 (in press)

CONFERENCE PROCEEDINGS

Quaglio M., Bezzo F., Gavriilidis A., Cao E., Galvanin F., A MINLP approach to model-based data mining for the quick development of nonlinear dynamic models, *Proceedings of the 2016 AIChE Annual Meeting*, 2016

Quaglio M., Fraga E. S., Galvanin F., Model-based design of experiments under structural model uncertainty, *Proceedings of the 27th European Symposium on Computer Aided Process Engineering*, 2017, pp. 145-150

Quaglio M., Fraga E. S., Galvanin F., Constrained model-based design of experiments

for the identification of approximated models, *Proceedings of the 18th IFAC Symposium on System Identification* 2018, pp. 515-520

Quaglio M., Waldron C., Pankajakshan A., Gavriilidis A., Galvanin F., A model-based data mining approach for outlier detection in kinetic modelling studies, *Proceedings of the 23rd International Conference on Chemical Reactors CHEMREACTOR-23*, 2018

Quaglio M., Fraga E. S. and Galvanin F., Statistical diagnosis of process-model mismatch by means of the Lagrange Multipliers test, *Proceedings of the 29th European Symposium on Computer Aided Process Engineering*, 2019, pp. 679-684

Quaglio M., Fraga E. S., Galvanin F., The evolution of approximated kinetic model structures, *Proceedings of the 2019 AIChE Annual Meeting*, 2019

CONFERENCE ABSTRACTS

Quaglio M., Bezzo F., Gavriilidis A., Cao E., Al-Rifai N., Galvanin F., Identification of kinetic models of methanol oxidation on silver in the presence of uncertain catalyst behaviour, *UK Catalysis Conference*, 2017

Quaglio M., Fraga E. S., Galvanin F., Model-based design of experiments for parameter precision under structural model uncertainty, *PSE@ResearchDay*, 2017

Quaglio M., Fraga E. S., Cao E., Gavriilidis A., Galvanin F., A model-based data mining approach for determining the domain of validity of approximated models, *ChemEngDayUK*, 2018

Quaglio M., Fraga E. S., Galvanin F., An evolutionary approach to kinetic modelling inspired by Lamarckian inheritance, *Workshop on Machine Learning and AI in (Bio)chemical Engineering*, 2019

Date:

Signature:.....

Abstract

Physics-based kinetic models are regarded as key tools for supporting the design and control of chemical processes and for understanding which degrees of freedom ultimately determine the observed behaviour of chemical systems. These models are formulated as sets of differential and algebraic equations where many state variables and parameters may be involved. Nonetheless, the translation of the available experimental evidence into an appropriate set of model equations is a time and resource intensive task that significantly relies on the presence of experienced scientists.

Automated reactor platforms are increasingly being applied in research laboratories to generate large amounts of kinetic data with minimum human intervention. However, in most cases, these platforms do not implement software for the online identification of physics-based kinetic models. While automated reactor technologies have significantly improved the efficiency in the data collection process, the analysis of the data for modelling purposes still represents a tedious process that is mainly carried out a-posteriori by the scientist.

This project focuses on how to systematically solve some relevant problems in kinetic modelling studies that would normally require the intervention of experienced modellers to be addressed. Specifically, the following challenges are considered: *i*) the selection of a robust model parametrisation to reduce the chance of numerical failures in the course of the model identification process; *ii*) the experimental design and parameter estimation problems in conditions of structural model uncertainty; *iii*) the improvement of approximated models embracing the available experimental evidence.

The work presented in this Thesis paves the way towards fully automated kinetic modelling platforms through the development of intelligent algorithms for experimental design and model building under system uncertainty. The project aims at the definition of comprehensive and systematic modelling frameworks to make the modelling activity more efficient and less sensitive to human error and bias.

Acknowledgements

This research project was supported by the 2016 H. Walter Stern Scholarship. I want to express my gratitude to the Centre for Nature-Inspired Engineering (CNIE) and to the UCL Chemical Engineering Department for this prestigious award and for the extraordinary opportunities of personal and professional development associated with it.

The work presented in this Thesis would have not been possible without the intellectual and emotional support of other people, to whom I am extremely grateful. I would like to express extreme gratitude to my advisor Dr Federico Galvanin for his wise guidance through the project and for his technical and emotional support. I also thank him for encouraging me to undertake a PhD degree and for always supporting me in the exploration of my ideas. I am extremely grateful to my advisor Prof Eric Fraga, for his consistent involvement in the project, his invaluable technical insights and his thoughtful advice.

I want to express great appreciation to Prof Asterios Gavriilidis and his research group. In particular, I must thank Enhong for providing me with the experimental data on methanol oxidation that I analysed during my Master's, which ultimately provided the initial inspiration for my PhD research project. I must also thank my friend Conor, who gave me the extraordinary opportunity to validate some of my model identification algorithms on the automated reactor system that he developed during his PhD. I must also acknowledge the CAPE lab, particularly Arun, Chunbing, Giannis, Panos and Harry for the fruitful discussions on modelling and data science, and for their constant encouragements and support.

I must thank my parents Carla and Paolo, my brothers Matteo and Simone, and my grandmother Paola for their constant emotional support through this journey despite the geographical distance. I must thank Conor and Niamh for being not only great flatmates, but also loyal friends. An acknowledgement is also due to all the other awesome and always supportive people I have had the pleasure to meet in London in these last 4 years: Laura, Anand, Ilaria, Zeynep, Mohammad, Andres, Sergio, Henry... They all contributed to make this experience worthwhile and unforgettable.

Impact statement

A number of novel techniques for the identification and improvement of physics-based kinetic models are proposed in this Thesis to reduce the time and resources required in kinetic modelling studies. The ultimate aim of this research project is the development of intelligent algorithms capable of identifying physics-based kinetic models without scientist supervision. This project was primarily motivated by the need for more systematic approaches to address modelling challenges in the context of Process Systems Engineering. Nonetheless, several impact areas are expected to benefit from this research, including areas outside the chemical engineering field.

Process improvement. Modern process industry is required to satisfy increasingly stricter constraints on environmental impact while ensuring profitability margins. To meet these requirements, the Quality by Design (QbD) paradigm dictates the necessity for detailed physics-based process models for an optimal design and operation of industrial plants (Yu, 2008).

Catalyst design. Catalysts are recognised as fundamental materials for the transition towards a green process industry. Nonetheless, the design and production of appropriate catalytic materials is an extremely challenging task, which relies on an efficient characterisation of process kinetics in the presence of a high number of different catalyst formulations (Thybaut et al., 2011).

Drug discovery. Antimicrobial resistance is now recognised as a fundamental threat to humanity (United Nations, 2016). Fast discovery of drugs and drug cocktails is an aspect of paramount importance to outpace the rate at which bacteria are developing drug resistance. This relies on a fast identification of PharmacoDynamic (PD) models from time-kill data to assess the potency of a given treatment in neutralising a certain bacterial species (Foerster et al., 2016).

Clinical practice. Accurate PharmacoKinetic (PK) models represent invaluable tools for the design of safe, non-invasive and effective clinical trials. Nevertheless, the identification of accurate PK models relies on an efficient extraction of information from small datasets to minimise the distress caused to the subject (Abbiati et al., 2018).

Algorithms sprouting from this research are expected to promote a faster identification of process models and a more rapid integration of QbD principles in existing and future process plants. Intelligent algorithms for kinetic modelling will also be applied for the fast characterisation of kinetics in the presence of novel formulations of catalytic materials. This will enable a faster deployment of effective catalysts in industrial processes and promote a faster transition towards an environment-friendly process industry. In pharmacology, smart algorithms for kinetic modelling will enable the rapid identification of PD models for the quick discovery of active pharmaceutical ingredients, drugs and drug cocktails. Ultimately, the rapid identification of PK models enabled by smart computational tools for kinetic modelling will contribute to increasing the understanding of human physiology. Particularly, such PK models may advocate the design of effective and less invasive clinical trials to diagnose and treat aggressive diseases, whose complex interactions with the human body are yet to be fully understood.

Contents

1 Introduction **21**

1.1 Motivation of the project 21

1.2 Open challenges in the automation of kinetic modelling 24

1.2.1 Online kinetic model identification 24

1.2.2 Robustness of model identification algorithms 27

1.2.3 Cognitive limits of available model identification algorithms 33

1.3 Contribution and structure of this Thesis 36

1.4 Computational resources 39

2 Literature survey **43**

2.1 Deterministic models 43

2.2 Model classes 44

2.3 Approaches for model structure building 46

2.4 Statistical model building and identification 48

2.4.1 Bridging modelling and experimental activity 48

2.4.2 Identifiability analysis 50

2.4.3 The parameter estimation problem 53

2.4.4 Statistical tools for model validation 54

2.4.5 Model-based design of experiments for model discrimination 58

2.4.6 Model-based design of experiments for parameter precision 60

2.5 Practical identifiability and model sloppiness 61

2.6 Robust regression and outlier detection 65

2.7 Model structure improvement 69

2.8 Summary of literature review 72

3 Online model reparametrisation for robust parameter estimation 75

3.1 Introduction . . . 75

3.2 Proposed methodology . . . 76

3.2.1 Primary parameter estimation . . . 78

3.2.2 Parametrisation update . . . 79

3.2.3 Secondary parameter estimation . . . 80

3.2.4 Optimal MBDoE for parameter precision . . . 81

3.3 Case study . . . 83

3.3.1 Automated model identification platform . . . 83

3.3.2 Modelling assumptions . . . 85

3.3.3 Objective and methods . . . 86

3.4 Results . . . 88

3.4.1 Simulated case: samples generated in-silico . . . 88

3.4.2 Real case: samples collected from the experimental platform . . . 91

3.4.3 Results discussion . . . 94

3.4.4 Computational times and problem size . . . 95

3.5 Final remarks . . . 97

4 Parameter estimation under structural model uncertainty 99

4.1 Introduction . . . 99

4.2 Proposed methodology . . . 100

4.2.1 Model-Based Data Mining for Parameter Estimation . . . 101

4.2.2 Support Vector Machine training . . . 103

4.2.3 Constrained Model-Based Design of Experiments . . . 105

4.3 Case studies . . . 106

4.3.1 Case study 1: ethanol dehydrogenation on copper . . . 106

4.3.2 Case study 2: methanol oxidation on silver . . . 114

4.3.3 Computational times and problem size . . . 121

4.4 Final remarks . . . 122

5 Diagnosis of model misspecification 127

5.1 Introduction . . . 127

5.2 Proposed methodology . . . 128

5.2.1 Parameter estimation . . . 129
5.2.2 Goodness-of-fit test . . . 130
5.2.3 Lagrange multipliers test . . . 130
5.3 Case studies . . . 134
5.3.1 Case study 1: baker's yeast growth model . . . 135
5.3.2 Case study 2: glucose-insulin interaction model . . . 138
5.3.3 Computational times and problem size . . . 143
5.4 Final remarks . . . 144

6 Evolution of kinetic model structures 147
6.1 Introduction . . . 147
6.2 Proposed methodology . . . 148
6.2.1 Diagnosis of model misspecification . . . 149
6.2.2 Identification of relevant effects . . . 150
6.2.3 Model evolution . . . 152
6.3 Case studies . . . 153
6.3.1 Case study 1: baker's yeast growth model . . . 153
6.3.2 Case study 2: glucose-insulin interaction model . . . 157
6.3.3 Results discussion . . . 162
6.3.4 Computational times and problem size . . . 164
6.4 Limitations of the ERI-based approach . . . 165
6.5 Final remarks . . . 166

7 Conclusion and future perspectives 169

Bibliography 174

Appendices 201

A Online RP - Simulated case: additional information 201

B Online RP - Additional simulated cases 203

C Online RP - Real case: additional information 205

D Ethanol dehydrogenation - Experimental data generated in-silico 209

E Methanol oxidation - Experimental data **211**

F Baker's yeast system - Experimental data generated in-silico **213**

G Glucose-insulin interaction system - Experimental data generated in-silico **215**

H Multivariate MMI-based analysis **219**

H.1 Lagrange multiplier test 219

H.2 Case study and results 221

H.3 On the computability of the multivariate MMI 223

List of symbols

Latin symbols

A	pre-exponential factor
b	offset of reliability map
b_i	adsorption coefficient of species i
c	hyperparameter associated with the robust estimator $\hat{\theta}_{DM}$
c_{ij}	correlation coefficient between θ_i and θ_j
C_i	concentration of species i
C_i^{IN}	concentration of species i at the inlet
C_i^{OUT}	concentration of species i at the outlet
d	scaling factor of parameter space (> 0)
E	set of candidate effects
E_a	activation energy
F	volumetric flowrate
g	generic scalar function
G	glucose concentration in plasma
G_b	basal glucose concentration
H_0	null hypothesis
H_a	alternative hypothesis
I	insulin concentration in plasma
$I(\varphi)$	map of model reliability
k	kinetic constant
K	kernel function
K_{eq}	equilibrium constant
$\mathscr{L}$	log-likelihood function

$\mathscr{L}_d$	log-likelihood function under parametrisation θ_d
$\mathscr{L}_e$	log-likelihood function under parametrisation θ_e
$\mathscr{L}_{DM}$	weighted log-likelihood function for robust regression
$\mathscr{L}_m$	log-likelihood function under parametrisation θ_m
M_i	generic model structure
M_e	evolved model structure
$\dot{n}_i$	molar flowrate of species i
N	number of samples in the available dataset Y
N_C	number of species considered in the model
N_e	number of candidate effects
N_f	number of functions in a given kinetic model
N^{MAX}	maximum number of samples collectable
N_R	number of reactions considered in the model
N_s	number of constraints in the parameter estimation problem
N_{sp}	number of samples to be designed
N_{sp}^{MAX}	maximum number of simultaneously designed samples
N_u	number of independent inputs in a given kinetic model
N_x	number of state variables in a given kinetic model
N_y	number of output variables in a given kinetic model
N_θ	number of non-measurable parameters in a given model
$p(\cdot)$	probability distribution
P_i	partial pressure of species i in the mixture
P_{TOT}	total pressure
r	biomass growth rate
r_j	rate of reaction j
R	ideal gas constant
t	time
$t(\alpha)$	t-value at significance α
t_{ref}	t-value
T	temperature
u_i	i-th model input in $\mathbf{u} \in U$
U	vector space of model inputs

v	flow velocity along the axial coordinate of microchannel
w	catalyst mass
$v_{\theta,ij}$	ij-th element of the covariance matrix $\mathbf{V}_\theta$
x_i	i-th system state in $\mathbf{x}$
X	insulin action associated with remote insulin receptor
Y	dataset available for model identification
Y'	reduced dataset $Y' \subseteq Y$
z	axial coordinate of microchannel
$z_{\alpha/2}$	two-tailed score of standard normal distribution with significance α

Matrices and vectors

$[\cdots]^T$	transpose of vector $[\cdots]$ [v.d. $\times$ 1]
$\mathbf{0}$	column array whose entries are all equal to 0 [v.d. $\times$ 1]
$\mathbf{1}_\theta$	column array whose entries are all equal to 1 [$N_\theta \times 1$]
$\mathbf{a}_k$	k-th order time derivative of $\hat{\mathbf{y}}$ at $t = 0$ [$N_y \times 1$]
$\mathbf{f}$	column array of functions [$N_f \times 1$]
$\mathbf{G}$	linear transformation of parameter space $\Omega \to \Theta$ [$N_\theta \times N_\theta$]
$\mathbf{G}_\mathrm{P}$	primary transformation of parameter space $\Omega \to \Theta$ [$N_\theta \times N_\theta$]
$\mathbf{G}_\mathrm{S}$	secondary transformation of parameter space $\Omega \to \Theta$ [$N_\theta \times N_\theta$]
$\mathbf{h}$	column array of functions [$N_y \times 1$]
$\mathbf{H}$	observed Fisher information matrix [$N_\theta \times N_\theta$]
$\hat{\mathbf{H}}_k$	information matrix associated with the k-th sample to be designed [$N_\theta \times N_\theta$]
$\mathbf{H}_d$	information matrix associated with parametrisation θ_d [$N \times N$]
$\mathbf{H}_e$	information matrix associated with parametrisation θ_e [$N_\theta + 1 \times N_\theta + 1$]
$\mathbf{H}_m$	information matrix associated with parametrisation θ_m [$N + N_\theta - 1 \times N + N_\theta - 1$]
$\mathbf{I}$	identity matrix [$N_y \times N_y$]
$\mathbf{I}_\theta$	identity matrix [$N_\theta \times N_\theta$]
$\mathbf{R}$	matrix of rotation of parameter space [$N_\theta \times N_\theta$]
$\mathbf{s}$	column array of functions of likelihood parameters [$N_s \times 1$]
$\mathbf{u}$	column array of independent control variables (model inputs) [$N_u \times 1$]
$\mathbf{U}$	right normalised eigenbasis of $\mathbf{H}$ [$N_\theta \times N_\theta$]

$\mathbf{V}_\theta$	covariance of parameter estimates in Θ $[N_\theta \times N_\theta]$
$\mathbf{V}_\omega$	covariance of parameter estimates in Ω $[N_\theta \times N_\theta]$
$\hat{\mathbf{V}}_\omega$	predicted covariance of parameter estimates in Ω $[N_\theta \times N_\theta]$
$\mathbf{W}_y$	weighting matrix $[N_y \times N_y]$
$\mathbf{W}_\theta$	weighting matrix $[N_\theta \times N_\theta]$
$\mathbf{x}$	column array of state variables $[N_x \times 1]$
$\mathbf{y}$	sample - column array of measured output variables $[N_y \times 1]$
$\mathbf{y}_i$	i-th sample in dataset Y $[N_y \times 1]$
$\hat{\mathbf{y}}$	column array of predicted output variables $[N_y \times 1]$
$\hat{\mathbf{y}}_i$	column array of predicted output variables for sample $\mathbf{y}_i$ $[N_y \times 1]$
α	column array of Lagrange multipliers $[\text{v.d.} \times 1]$
$\hat{\alpha}$	column array of estimates for Lagrange multipliers $[\text{v.d.} \times 1]$
θ	column vector of parameters in parameter space Θ $[N_\theta \times 1]$
θ^*	column vector of target parameters in parameter space Θ $[N_\theta \times 1]$
$\hat{\theta}$	maximum likelihood estimate for $\theta \in \Theta$ $[N_\theta \times 1]$
θ_d	column vector of parameters in diagnostic parameter space $[N \times 1]$
$\hat{\theta}_d$	constrained maximum likelihood estimate for θ_d $[N \times 1]$
$\hat{\theta}_{DM}$	robust ML estimate for $\theta \in \Theta$ obtained maximising $\mathscr{L}_{DM}$
θ_e	extended vector of parameters for testing effect relevance $[N+1 \times 1]$
$\hat{\theta}_e$	constrained maximum likelihood estimate for θ_e $[N+1 \times 1]$
$\hat{\theta}_{LMS}$	robust *LMS*-estimate of $\theta \in \Theta$ $[N_\theta \times 1]$
θ_m	extended vector of parameters in diagnostic parameter space $[N+N_\theta-1 \times 1]$
$\hat{\theta}_m$	constrained maximum likelihood estimate for θ_m $[N+N_\theta-1 \times 1]$
$\hat{\theta}_M$	robust *M*-estimate of $\theta \in \Theta$ $[N_\theta \times 1]$
$\hat{\theta}_s$	constrained maximum likelihood estimate of $\theta \in \Theta$ $[N_\theta \times 1]$
Λ	diagonal matrix whose ii-th element is λ_i $[N_\theta \times N_\theta]$
Σ_y	covariance of measurement error for sample $\mathbf{y}$ $[N_y \times N_y]$
φ_i	vector of experimental conditions associated with sample $\mathbf{y}_i$
φ_i^*	optimised experimental conditions of sample $\mathbf{y}_i$
ω	column vector of parameters in parameter space Ω $[N_\theta \times 1]$
$\hat{\omega}_\mathrm{P}$	column vector of parameter estimates computed with $\mathbf{G} = \mathbf{G}_\mathrm{P}$ $[N_\theta \times 1]$
$\hat{\omega}_\mathrm{S}$	column vector of parameter estimates computed with $\mathbf{G} = \mathbf{G}_\mathrm{S}$ $[N_\theta \times 1]$

Greek symbols

α	statistical significance
α_i	Lagrange multiplier
β_i	binary variable associated to the i-th sample in $\mathscr{L}_{DM}$
$\hat{\beta}_i$	estimate for the binary variable β_i
γ	decay length of Gaussian radial basis function
ε	small positive constant
η_i	i-th effect in E
θ_i	i-th model parameter
$\hat{\theta}_i$	estimate for the i-th model parameter
Θ	original vector space of model parameters
κ	condition number
λ_i	i-th eigenvalue of $\mathbf{H}$
ν_{ij}	stoichiometric coefficient of the i-th species in the j-th reaction
ξ_d	Lagrange multipliers statistic computed with parametrisation θ_d
ξ_e	Lagrange multipliers statistic computed with parametrisation θ_e
ξ_{LM}	generic Lagrange multipliers statistic
ξ_{LR}	likelihood ratio statistic
ξ_m	Lagrange multipliers statistic computed with parametrisation θ_m
ξ_W	Wald statistic
ρ	generic function
τ	time horizon
Φ	space of experimental conditions for a sample $\mathbf{y}$
χ^2	generic χ^2 statistic
$\chi^2_k(\alpha)$	χ^2 value at significance α and degree of freedom k
χ^2_{ref}	95% value computed from a χ^2 distribution
χ^2_Y	sum of normalised squared residuals associated with the fitting of dataset Y
ψ	experimental design metric
Ω	transformed vector space of model parameters
∇	gradient operator in parameter space

Acronyms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CAPE	Computer-Aided Process Engineering
ED	Experimental Design
ERI	Effect Relevance Index
HPLC	High-Performance Liquid Chromatograph
IDR	Insulin Delivery Rate
IVGTT	IntraVenous Glucose Tolerance Test
LHHW	Langmuir-Hinshelwood-Hougen-Watson
LMS	Least Median of Square
MAP	Maximum A Posteriori
MBDM	Model-Based Data Mining
MBDoE	Model-Based Design of Experiments
MIMO	Multiple Input Multiple Output
MINLP	Mixed-Integer NonLinear Program
MMI	Model Modification Index
ML	Maximum Likelihood
ODE	Ordinary Differential Equation
PSE	Process Systems Engineering
RG	Regularisation
RP	Reparametrisation
SCA	Surrogate Cognitive Agent
SLSQP	Sequential Least Squares Quadratic Program
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Motivation of the project

The reduction in the cost of computational power over the last decades led to an exponential increase in the employment of computational methods to address engineering challenges. In particular, the field of process systems engineering is thriving in the digital revolution (Sargent, 2005). The sub-field of Computer-Aided Process Engineering (CAPE) stemmed directly from the increasing access to cheap computational resources and it is transforming dramatically the way industrial processes are developed, designed and optimised. The CAPE approach involves the implementation of process models into computer programs with the aim of increasing process understanding and identify optimal process design and control solutions through simulations and numerical analyses of process data (Sargent, 1967).

As a consequence of the successful marriage between Process Systems Engineering (PSE) and computer science, the modelling of chemical and biochemical kinetics is becoming an increasingly relevant research topic both in industry and academia (Bonvin et al., 2016). This trend can be observed in Figure 1.1, which reports the number of scientific contributions on kinetic modelling that were published in Engineering Journals in the last 80 years (Web of Science, 2019). In particular, *phenomenological* kinetic models are regarded as extremely valuable tools in CAPE. These models are typically formulated as systems of differential and algebraic equations whose mathematical structure reflects the causal mechanisms of the physical system. Accurate phenomenological models are recognised as important means for understanding which degrees of freedom ultimately determine the behaviour of chemical processes (Rosenblueth and Wiener, 1945). Nevertheless, simple mathematical descriptions of dynamic phenomena represent invaluable tools for supporting

non-empirical process design and optimisation (Biegler et al., 1997). In fact, models derived from appropriate physics-based hypotheses can be used to predict the dynamic behaviour of physical systems also at conditions that were not previously observed (Hancock and Compton, 1999). Accurate phenomenological models may be employed to identify non-trivial design and control solutions to minimise the environmental impact of chemical processes while respecting constraints on process profitability.

Despite the priceless contribution of many researchers in the fields of model building and design of experiments, the phenomenological modelling of kinetic phenomena remains a time and resource intensive task that significantly relies on the intuition of experienced modellers and experimentalists. The kinetic model building process can be summarised in three fundamental steps (Walter and Pronzato, 1997):

1. *Model formulation.* The prior knowledge and the experimental evidence available on the system behaviour are distilled into an opportune set of modelling hypotheses. These hypotheses are then translated mathematically into a set of kinetic model equations.

2. *Parameter estimation.* The kinetic parameters involved in the model structure must be precisely estimated by fitting experimental data.

3. *Model validation.* Statistical tools are employed to validate the modelling hypotheses against experimental observations.

Experimental data are typically required in all the aforementioned stages, especially if there is a significant lack of prior knowledge on the kinetic behaviour of the system. Hence kinetic studies may require extensive amounts of time and resources both for performing experiments and for analysing experimental data.

Significant effort has been devoted by the scientific community to the mitigation of the experimental and analytical burden required to identify and validate kinetic models. Important steps towards the reduction in the cost of kinetic studies are 1) the coupling of automated, small-scale flow reactor technologies with online analysis equipment for the quick collection of experimental data (Goodell et al., 2009) and 2) the employment of advanced statistical tools for planning informative experiments with the aim of minimising the cost, time and amount of resources required for the experimentation (Asprey and Macchietto, 2000). Specifically, a variety of Model-Based Design of Experiments (MBDoE)

techniques have been proposed in the literature to plan optimally informative experiments for achieving specific objectives, e.g. selecting the best model out of a set of candidates (Hunter and Reiner, 1965; Buzzi-Ferraris et al., 1984) and/or improving parameter precision in an already selected model (Zullo, 1991; Prasad and Vlachos, 2008; Chakrabarty et al., 2013; Galvanin et al., 2013; Stamati et al., 2016). Some recent works in the literature have also considered to couple automated flow reactor technologies with advanced modelling algorithms with the aim of developing fully automated platforms for the identification of phenomenological kinetic models (McMullen and Jensen, 2011; Bournazou et al., 2016; Echtermeyer et al., 2017; Waldron et al., 2019b). In these platforms, the model is identified by a numerical algorithm in the course of an unmanned experimental campaign. As soon as new samples are collected from the system, these are analysed online by the software to improve process understanding. Optimal experimental conditions for model identification are then computed by an MBDoE routine and transmitted to the automated reactor to collect additional samples and complete the model identification process.

Such platforms have the potential of dramatically speeding up the modelling of kinetic phenomena and, consequently, the discovery and the study of new chemical processes. Nonetheless, there still remain a significant number of computational challenges that need to be addressed to promote their diffusion in research laboratories. The computational limits of state-of-the-art model identification platforms are associated primarily with aspects of the modelling activity that are currently difficult to automate, e.g. *i*) the formulation of an appropriate set of modelling hypotheses and their translation into a set of kinetic model equations, *ii*) the estimation of parameters in the presence of approximated kinetic model structures, *iii*) the design of optimal experiments to improve parameter precision in the presence of approximated models, *iv*) the refinement of the modelling assumptions embracing the available experimental evidence and *v*) the solution of parameter estimation problems in the presence of high parameter correlation and/or low parameter sensitivity.

The identification of systematic techniques to address the aforementioned challenges is the objective of this research project. The work presented in this Thesis aims at developing a robust framework for integrating modelling and experimental activities decoupling and targeting the various sources of uncertainty involved in the study of kinetic phenomena. The final aim of the project is the definition of a comprehensive and systematic approach to the identification of kinetic models in the attempt of making the modelling activity more

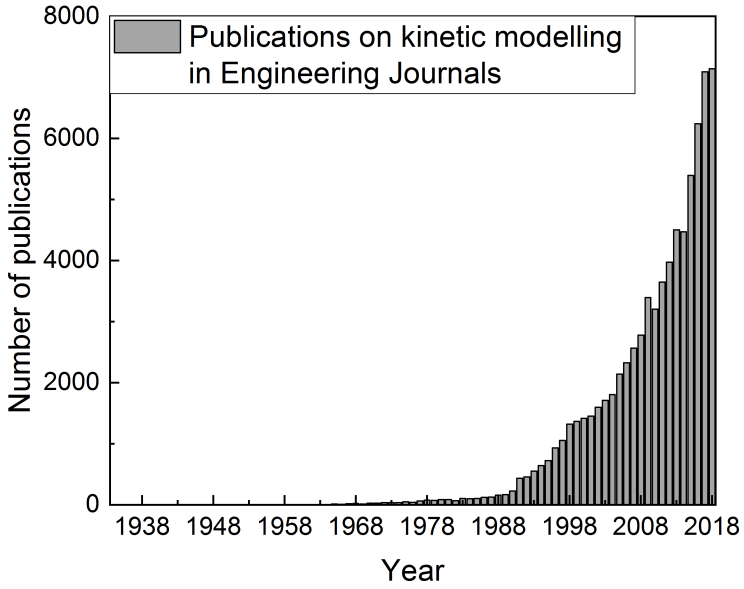


Figure 1.1: Number of publications per year on the topic of *kinetic modelling* in Engineering Journals (Web of Science, 2019).

efficient and less sensitive to human bias.

1.2 Open challenges in the automation of kinetic modelling

In this section, the main challenges associated with the building, identification and improvement of kinetic models are introduced and discussed. It is shown that the the definition of systematic approaches to address these challenges is key to increase the efficiency of kinetic modelling studies and advance the state-of-the-art of automated kinetic model identification platforms.

1.2.1 Online kinetic model identification

Advances in the fields of automation and the development of fast measurement devices enabled the coupling of automated reactors with fast-response sensors for online monitoring (Goodell et al., 2009). The presence of online measurement systems in these devices allows the generation of high frequency kinetic data during reactor operation with minimum human intervention. In particular, when small-scale flow reactor technologies are employed, these automated platforms enable the automatic generation of substantial amount of kinetic information consuming small amounts of materials and concomitantly reducing the cost and the risks associated with the experimentation (Jeraal et al., 2018).

Automated flow reactors have been employed in a wide variety of situations from process monitoring (Malig et al., 2017) to screening of operating conditions (Walsh et al.,

2005; Gromski et al., 2019) and reaction discovery (Steiner et al., 2019). Automated flow reactors were also successfully coupled to algorithms for online sequential design of experiments (McMullen and Jensen, 2010; Moore and Jensen, 2012; Fabry et al., 2014; Holmes et al., 2016). After every experiment is terminated and new data are collected by these platforms, an algorithm constructs a black-box representation of the physical system, e.g. a response surface (Box and Lucas, 1959) for designing the following experiment with the aim of optimising the reaction performance (e.g. the conversion or the yield). These works demonstrated that automated reactor platforms can be responsive and adapt their actions to the behaviour of the system. However, these automated devices do not exploit the collected data for the online development and identification of *physics-based* models. A major consequence of this is that optimised reaction conditions identified through a black-box approach in the lab-scale equipment are not necessarily transferable to the design, optimisation and control of equipment at the industrial scale.

Only few works are available in the literature in which algorithms for the online identification of phenomenological models were coupled to automated reactor systems (McMullen and Jensen, 2011; Bournazou et al., 2016; Echtermeyer et al., 2017; Waldron et al., 2019b). In these works, numerical routines for parameter estimation and optimal Model-Based Design of Experiments (MBDoE) were employed online to drive the experimental campaigns with the aim of selecting the best model among a set of given phenomenological models (i.e., model discrimination) (McMullen and Jensen, 2011) and/or improving the statistical quality of the parameter estimates for a given model structure (McMullen and Jensen, 2011; Bournazou et al., 2016; Echtermeyer et al., 2017; Waldron et al., 2019b). The further diffusion of these promising systems in research laboratories is hampered by a number of limitations that are present in state-of-the-art model identification algorithms. These limitations are associated primarily with aspects of the modelling activity that are currently complex to automate. Some of these aspects are introduced in the following list and further discussed in the following sections.

1. *Selection of the modelling hypotheses.* Practical rules for constructing kinetic models have been proposed in the literature (Fogler, 2005). However, these rules are usually complex to formalise and/or generalise. Automated approaches for the generation of kinetic models are also available (Oliveira et al., 2016). Nonetheless, the application of automated approaches for kinetic model construction typically results in the de-

velopment of computationally intractable models that are impractical for engineering purposes. Hence, in most cases, an appropriate set of simplifying modelling hypotheses has to be manually selected by the scientist and provided as an input to model identification algorithms in the form of kinetic model equations.

2. *Parameter estimation in the presence of structural model uncertainty.* The selection of the modelling hypotheses is a process that relies almost entirely on the decisions of thoughtful researchers. Thus, being sensitive to human bias, hypotheses selection shall be treated as an additional source of uncertainty in phenomenological modelling (Galvanin et al., 2012). If the proposed model structure is misspecified or excessively approximated, it may not represent accurately the distribution of the data across the entire range of explorable experimental conditions. Only data collected at conditions where the modelling assumptions are *valid* should be considered for parameter fitting (Tsay et al., 2017). Nevertheless, the range of conditions in which the modelling assumptions can be considered accurate is normally not known a-priori and has to be learnt through experimentation. Only few works have considered the possibility of systematically quantifying the domain of reliability of approximated models (Kahrs and Marquardt, 2007). Nonetheless, in most cases the domain of reliability of kinetic models is only qualitatively inferred by the researcher though the formulation of conjectures on the actual behaviour of the system.

3. *Refinement of the modelling hypotheses.* If the postulated model structure is falsified by experimental observations, a reformulation of the modelling hypotheses may be required. The modelling hypotheses should be improved embracing the available experimental evidence and the model structure should be changed accordingly. Nevertheless, improving an approximated model structure maintaining its physical significance is a task that relies almost entirely on human intuition and experience.

4. *Solution of ill-conditioned parameter estimation problems.* Parameter estimation and MBDoE problems are normally recast as optimisation problems and solved numerically. Both optimisation problems may be ill-conditioned (Chiş et al., 2014; Wilson et al., 2015; White et al., 2016). More specifically, parameter estimation problems may not admit a unique solution and objective functions considered in MBDoE problems may be undefined (e.g. because of a division by zero). These identifiability

problems may be the consequence of a poor choice of the model parametrisation and/or the availability of a poorly informative dataset (Söderström and Stoica, 1989). A number of strategies have been proposed in the literature to diagnose and address model identifiability problems. However, these approaches are typically case dependent and the intervention of a scientist is normally required to reformulate the model identification problem.

5. *Recognition of irrelevant data for kinetic modelling.* The execution of kinetic experiments requires the stimulation of a physical system in a controlled environment through the manipulation of input variables and the detection of the system response through the measurement of some system state. However, in any setup there are control limits and it may be impossible to completely eliminate external disturbances. The fitting of data collected in the presence of significant system disturbances may invalidate the model identification process (Hampel, 1985). These data should be recognised and neglected for kinetic modelling purposes (Özyurt and Pike, 2004). Several approaches were proposed to recognise the presence of disturbances in the contexts of process monitoring and fault detection (Venkatasubramanian et al., 2003; Yin et al., 2014). These approaches may be classified as model-based and data-based. Nevertheless, the application of model-based approaches relies on a substantial confidence on the assumptions underlying the model, while the application of data-based approaches relies on the presence of a substantial amount of process data. None of these requirements is typically satisfied in kinetic modelling studies, where the resources available for experimentation are normally scarce and the behaviour of the system in the absence of disturbances is highly uncertain.

The improvement of available modelling techniques and the development of novel robust approaches for addressing the aforementioned challenges are important aspects for the design of future automated platforms for online kinetic modelling.

1.2.2 Robustness of model identification algorithms

The structure of a state-of-the-art platform for model identification is given in Figure 1.2. In the Figure, arrows represent flows of information, while lightning-shaped symbols represent sources of uncertainty. In state-of-the-art platforms for physics-based kinetic modelling, a set of possible model structures is provided as an input by the researcher. The model identification algorithm is then asked to select the best model structure among the

proposed ones and provide estimates for its kinetic parameters. To achieve this task, the model identification algorithm can perform experiments and collect samples from an automated experimental setup.

Being sensitive to human decisions, the proposed model structures may be inappropriate for system identification purposes. In fact, the model structures proposed by the scientist may be affected by identifiability issues associated with limits in the observability/controllability of the system. Furthermore, none of the proposed model structures may appropriate to model the process. More specifically, all the proposed structures may have been derived from inappropriate modelling hypotheses. In addition to these aspects, experimental disturbances may occur in the course of the unmanned experimental campaign. A failure of the model identification algorithm in handling these uncertainties may lead to the invalidation of the model identification process and to a significant waste of experimental resources.

The potential consequences on the modelling process that are associated with the aforementioned uncertainties are further discussed in the following subsections. It is argued that there is a significant need for robust computational tools capable of dealing autonomously with these sources of uncertainty to encourage the transition towards the automation of kinetic modelling studies, but also to promote the diffusion of good modelling practice in research laboratories.

1.2.2.1 Robustness towards model identifiability issues

Once a model structure is selected, its identification requires the estimation of its parameters by fitting experimental data. Due to observability and controllability constraints in the experimental setup, it may be impossible to perform an experiment to obtain the information required to estimate the model parameters (Saccomani et al., 1997). Identifiability tests may be conducted before any experiment is performed to check if it is possible to uniquely estimate the model parameters given the observability and controllability limits of the setup (Raue et al., 2014). Systematic approaches for conducting a priori identifiability analysis were proposed in the literature (Audoly et al., 2001; Sedoglavic, 2002; Saccomani et al., 2003), and computational tools for identifiability analysis are also available (Bellu et al., 2007; Chiş et al., 2011; Anguelova et al., 2012). Nonetheless, even if the model structure provided to the model identification algorithm satisfies the requirements for a priori identifiability, it may still be extremely challenging to retrieve its parameters using numerical

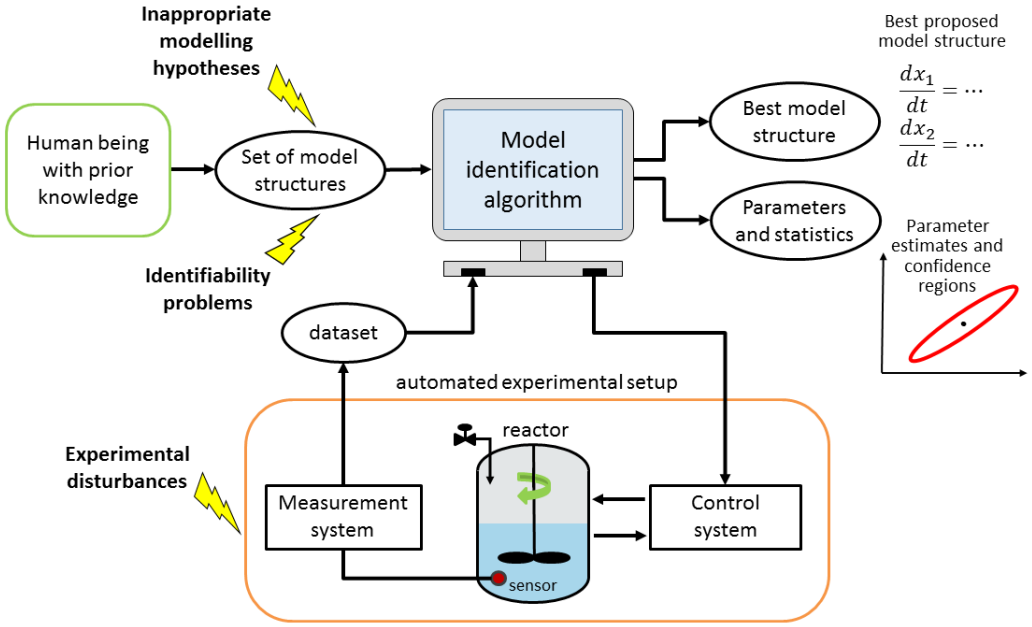


Figure 1.2: Simplified diagrams showing the flows of information in an automated kinetic model identification platform. Lightning-shaped symbols highlight sources of uncertainty in the model identification problem.

routines (Transtrum et al., 2010).

Whenever a model structure is provided as input, the model identification algorithm is required to solve sequentially *1*) a parameter estimation problem given the available dataset (Bard, 1974) and *2*) an experimental design problem to design following samples with the aim of minimising the uncertainty on the parameter estimates (Franceschini and Macchietto, 2008b). The solution of both problems may require the employment of numerical optimisation routines and their effectiveness requires their respective objective function to be well-conditioned (Wilson et al., 2015; White et al., 2016). Ill-conditioned objective functions derive from the attempt of identifying models whose parametrisation is sloppy given the available dataset (Chiş et al., 2014; White et al., 2016). Sloppiness arises when measured model responses are poorly sensitive to the change of some parameters and/or measurements do not carry sufficient information to bring parameter correlation below a critical threshold (typically considered as high as 95%). Whenever these circumstances occur, the eigenvalues of the covariance matrix of the parameter estimates span over a wide range of orders of magnitude, i.e., the condition number of the covariance matrix is very high (Higham, 1996).

Numerical failures may occur in the course of the model identification problem in the

presence of a sloppy parametrisation. These may be classified as follows:

- *False convergence.* Ill-conditioned objective functions both in the parameter estimation and in the optimal MBDoE problem may cause numerical optimisation routines to fail in converging to the optimal solution (Higham, 1996).

- *Inaccuracy in the computation of gradients.* The calculation of the sensitivities (i.e., partial derivatives in the parameter space) using direct differential methods is frequently impractical. As a consequence, numerical differentiation routines are regularly employed in model building practice (Saltelli et al., 2000). The numerical computation of sensitivities requires a perturbation of the model parameter values. The computed sensitivities are *sensitive* to the choice of the perturbation. In the presence of a sloppy parametrisation, the applied perturbation may not be appropriate to accurately quantify the gradient in the parameter space (Higham, 1996). As a consequence, the covariance matrix computed as a function of the parameter sensitivities may be inaccurate, affecting the model validation process and the design of following experiments (Pukelsheim, 2006).

- *Inaccuracy in the inversion of matrices.* In the presence of a sloppy parametrisation, the covariance matrix of the parameter estimates is ill-conditioned (White et al., 2016). The solution of an optimal MBDoE problem requires the inversion of an ill-conditioned covariance matrix if the parametrisation is sloppy (Franceschini and Macchietto, 2008b).

The optimisation of ill-posed functions may lead to significant numerical failures in the course of an unmanned experimental campaign with the concomitant waste of experimental resources. Improving the robustness of automated model identification platforms towards model sloppiness is key to further promote their employment in the discovery and study of kinetic phenomena.

1.2.2.2 Robustness towards structural model uncertainty

Mathematical models are never perfect descriptions of the underlying physical phenomenon (Box and Draper, 1987). This shall be regarded as a strength of mathematical modelling rather than a weakness (White et al., 2016). In fact, models built from a thoughtful selection of simplifying hypotheses provide insights on which are the fundamental degrees of freedom that are ultimately responsible for a certain system behaviour (Rosenblueth and

Wiener, 1945). However, the selection of the modelling hypotheses is still a process that relies almost entirely on the decisions of thoughtful researchers. Thus, being sensitive to human errors, hypotheses selection shall be treated as an additional source of uncertainty in mechanistic modelling.

Conventional parameter estimation and Model-Based Design of Experiments (MBDoE) techniques do not consider structural model uncertainty in the formulation of fitting cost functions and experimental design metrics (Asprey and Macchietto, 2000). Model identification approaches implemented in most modelling algorithms assume that the parameter estimates will converge to a *true* parameter value as the amount of fitted samples increases (Bard, 1974). In the presence of misspecified or approximated model structures, true parameter values may not exist and estimates typically do not converge. The estimation of non-converging parameters may result in significant numerical failures and waste of resources, especially if the model is nonlinear and its identification is performed online on an unsupervised experimental platform.

A further aspect associated with the identification of an approximated model is that the discrepancy between observations and model predictions is the consequence of both measurement error and process-model mismatch. If the model is approximated, it may be possible to accurately fit only data collected within the model validity domain, namely the range of conditions where the simplifying modelling hypotheses may be considered *valid*. The inclusion in the parameter estimation problem of data collected outside the model validity domain may result in the computation of estimates with questionable physical significance, a degradation of the model fitting quality and a loss of model predictive capability.

The experimental design stage in the parameter estimation process is also affected by the presence of structural model uncertainty. MBDoE methods for parameter precision properly account for the uncertainty that is intrinsically present in the measurement system (i.e., the measurement noise) and how this uncertainty propagates to the parameter estimates (Bard, 1974; Walter and Pronzato, 1997). However, standard MBDoE tools do not account for systematic errors nor for the uncertainty and approximation that may be present in the candidate model equations, i.e., they assume that the structure of the model used to perform experimental design is *exact*. As a consequence, the inconsiderate application of standard MBDoE methodologies in the presence of an approximated model structure may lead to the execution of experiments and collection of data outside the model validity domain. Only

few works in the literature have considered the extension of available MBDoE methods to account for structural model uncertainty and experimental disturbances (Galvanin et al., 2011, 2012).

Data collected at conditions where the model is not valid shall be regarded as irrelevant for the estimation of the model parameters. Nevertheless, the geometry of the domain of model validity is normally unknown a priori and has to be inferred through experimentation (Kahrs and Marquardt, 2007; Tsay et al., 2017). The development of robust estimators and MBDoE criteria embracing structural model uncertainty is one of the aims in this research project.

1.2.2.3 Robustness towards experimental disturbances

Kinetic experiments involve an interaction with the physical system through the control of certain input variables and the observation of the dynamic response in some other output variables. Nevertheless, the ability of the experimentalist or even an automated system to control an experimental setup is never perfect. Disturbances may occur in the course of the experimental campaign resulting in anomalies in the behaviour of the system. Examples of disturbances in reactor-based setups may be the presence of contaminants in the reactor feed, control offsets and leakages. Disturbances can never be completely eliminated and shall be taken into account in the model identification process (Özyurt and Pike, 2004). In fact, system disturbances can lead to the collection of outliers. Outliers are defined by Rousseeuw and Leroy (1987) as data that deviate from the assumptions and their inclusion in the dataset may have a dramatic impact on the modelling process (Huber, 1981).

The concept of breakdown point was developed to assess the sensitivity of a parameter estimator towards the presence of outliers in the dataset (Huber, 1964). Hampel defines the breakdown point as the smallest fraction of outlier contamination in the dataset that can carry the parameter estimates beyond any finite bound (Hampel, 1985). For traditional estimators based on maximum likelihood (e.g. the least squares method), the breakdown point approaches 0 as the number of fitted samples increases, meaning that one outlier in the dataset is sufficient to invalidate the parameter estimation. The acknowledgement of this weakness led to the formulation of alternative estimators that are insensitive to outlier contamination in the dataset and a whole subarea of statistics, namely the field of robust regression (Huber, 1964; Rousseeuw and Leroy, 1987; Özyurt and Pike, 2004). Nevertheless, parameter estimation is performed using standard estimators in most kinetic modelling

studies and outlier detection is performed using heuristic rules. As an example, a popular approach used to label *bad* samples is the method of the material balance, which consists of quantifying the discrepancy in the atom balances between the inlet and the outlet of the reactor (Galvanin et al., 2018; Waldron et al., 2019a). Kinetic experiments in which this discrepancy is above a certain threshold (typically chosen between 5% and 10%) are considered too inconsistent to be used for kinetic modelling and are removed from the dataset. A major drawback of the material balance method is that it is blind towards certain types of disturbances. For instance, samples collected in the presence of a significant temperature offset may not be detected as outliers because such disturbance may not affect the input-output balance of the atomic species.

The employment of robust estimators derived from sound statistical foundations should become common tools in kinetic modelling studies (Buzzi-Ferraris and Manenti, 2009). Especially in online kinetic modelling, a failure in recognising outliers could lead to the execution of a suboptimal experimental campaign and the ultimate failure of the modelling algorithm in selecting an appropriate model structure and estimating its parameters.

1.2.3 Cognitive limits of available model identification algorithms

A broad variety of problems in the field of process systems engineering cannot be solved effectively by human intuition only. The nature of these problems, from process design, optimisation and control, is frequently multi-objective, highly nonlinear, constrained by physics, manufacturing capabilities, legislation and costs. Information technology has already augmented human cognitive capabilities and helped improve chemical processes to a point that would have not been reachable solely with human intuition. This cognitive leap in process systems engineering was enabled primarily by the development of efficient numerical equation solvers (Brenan et al., 1987) and optimisation algorithms (Floudas and Pardalos, 2013). These algorithms have been extensively employed also in the field of kinetic modelling and represent the foundation of state-of-the-art software for process modelling and simulation, e.g. the general PROcess Modelling System gPROMS® (PSE gPROMS, 2017), MATLAB® Simulink (Mathworks MATLAB, 2015), the software EFCOSS (Environment for Combining Optimization and Simulation Software) (Rasch and Bücker, 2010) and the Engineering Equation Solver EES (F-Chart Software EES, 2017).

Once an opportune kinetic model structure is selected, parameter estimation and optimal experimental design problems can be recast as optimisation problems and solved using

numerical optimisation routines (Walter and Pronzato, 1997). However, the problem of formulating an appropriate set of kinetic model equations is multi-objective and involves substantially different challenges. The aim in physics-based kinetic modelling is to build a set of model equations such that *1*) the structure of the equations reflects the causal dynamic mechanisms of the physical system *2*) the parameters involved in the model can be uniquely retrieved by fitting experimental data *3*) the parameters can be precisely estimated given the available experimental budget and *4*) the structure is simple enough to be used for practical engineering purposes.

Some authors proposed to recast the problem of model structure formulation as an optimisation problem and solve it employing numerical optimisation routines (Cozad et al., 2015; Tsay et al., 2017; Wilson and Sahinidis, 2017; Neumann et al., 2019). Nonetheless, there may be infinite model structures that are accurate in representing the experimental observations. For such reason, model structure selection problems are inherently ill-conditioned and the model structure space has to be manually constrained in order to make the problem solvable. A limitation of this approach is that none of the model structures in the constrained structure space may be appropriate to model the physical phenomenon. Genetic programming was proposed as a mean of exploring effectively vast solution spaces (Banzhaf et al., 2015). Applications of genetic programming to structural equation modelling are also available in the literature (Florin Metenidis et al., 2004; Xiao-lei Yuan et al., 2008; Gandomi and Alavi, 2011). However, genetic approaches rely on the construction and identification of a substantial number of model structures. The estimation of parameters in a high number of kinetic models may be impractical, especially if the models are nonlinear in the parameters (Florin Metenidis et al., 2004; Transtrum et al., 2010) and affected by problems of identifiability.

A number of software packages have been developed for supporting the scientist in the construction of reaction networks starting from given chemistry rules, i.e., a potential sets of elementary reactions or reactions families that can occur in the system (Oliveira et al., 2016). Some of these computational tools are *NetGen* from Broadbelt et al. (1994), the *Reaction Mechanism Generator* developed by Song (2004), the *Reaction Modeling Suite* proposed by Katare et al. (2004) and the *Genesys* software developed by Vandewiele et al. (2012). This list is by no means comprehensive and more detailed overviews on software for the automated generation of kinetic models can be found in Ugi et al. (1993); Katare et al.

(2004); Klein et al. (2005); Van de Vijver et al. (2015a). Software for automated generation of reaction networks were employed on a number of different modelling problems, from the modelling of pyrolysis (Broadbelt et al., 1994; Van de Vijver et al., 2015b), hydrocracking (Mizan and Klein, 1999) and aromatisation (Bhan et al., 2005) to the description of syngas production (Seyedzadeh Khanshan and West, 2016) and biomass conversion (Rangarajan et al., 2010). A more comprehensive overview on the applications of automated model building software can be found in Oliveira et al. (2016).

These algorithms generate exhaustive reaction networks by accessing a user-defined library of possible chemical species and a database of elementary reactions. They then parse the reaction network into a set of kinetic model equations (Katare et al., 2004). This typically results in the formulation of models involving a high number of species and reactions, which may be inappropriate for parameter estimation and simulation purposes. The generated model may be simplified by removing irrelevant reactions from the full mechanism. However, information on which are the relevant reactions occurring in the system is seldom available and heuristic model reduction rules are used instead. These rules may involve limiting the maximum molecule size, excluding chemical species, ignoring reaction families and so on (Oliveira et al., 2016). The computational complexity associated with the application of such reaction mechanism generators makes them inappropriate for application in online model identification platforms. Furthermore, such algorithms require the availability of substantial prior information on the system in the form of species and elementary reactions libraries. Possible elementary reactions may be identified through *ab initio* quantum mechanical simulations (Lu and Yang, 2004). However, molecular simulations typically require substantial computational resources and numerical results of *ab initio* calculations are sensitive to a high number of user-defined assumptions on the system behaviour (Parr, 1980).

It is extremely challenging to formalise the problem of selecting and refining modelling hypotheses in a computer program and the introduction of human bias is currently unavoidable. This bias is provided to model identification algorithms in the form of candidate model structures derived from simplifying hypotheses. If none of the proposed structures is appropriate to describe the dynamics of the system, these should be *evolved* embracing the available experimental evidence. The concept of evolution in structural equation modelling is borrowed from the literature on genetic programming (Banzhaf et al., 2015) and

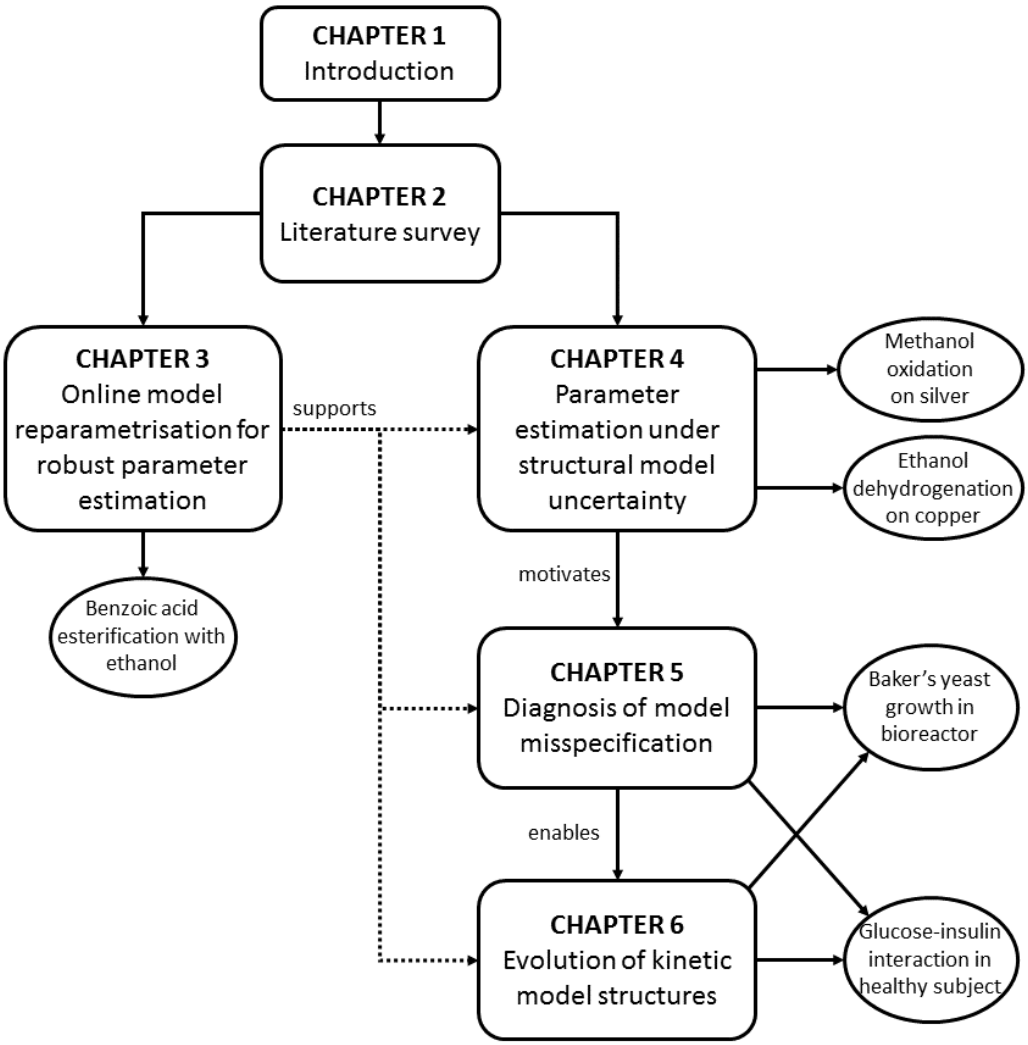


Figure 1.3: Roadmap representing the structure of this Thesis.

will be used to refer the act of improving a kinetic model by modifying its mathematical structure. Even for an experienced scientist, refining the modelling hypotheses is not a trivial task and no algorithm exists that can perform such operation efficiently and taking into account all the aforementioned modelling requirements. One of the aims of this research project is the definition of computationally tractable approaches to support the scientist in diagnosing model misspecification and evolving approximated model structures embracing both experimental evidence and prior knowledge available on the system.

1.3 Contribution and structure of this Thesis

In the previous sections, the main sources of uncertainty associated with the study of kinetic phenomena were illustrated and discussed. The challenge embraced in this research project

is the development of intelligent algorithms for kinetic modelling that are robust, i.e. insensitive, to the presence of these uncertainties. In particular, the work presented in this Thesis focuses on four fundamental problems encountered in kinetic modelling practice:

1. The online estimation of kinetic parameters in the presence of a sloppy parametrisation, i.e., in the presence of extreme parameter correlation and/or poor sensitivity of the measured model responses to a change in the parameter values.

2. The estimation of kinetic parameters and determination of the model validity domain (i.e., the range of conditions in which the modelling assumptions may be considered acceptable) in the presence of approximated kinetic model structures and in the context of online kinetic modelling.

3. The diagnosis of model misspecification in approximated model structures, namely the detection of the model components that are inappropriately specified and require reformulation.

4. The systematic improvement of approximated kinetic model structures embracing both prior knowledge and experimental evidence available on the dynamic behaviour of the system.

A roadmap representing the structure of this Thesis is given in Figure 1.3. This work is organised in seven Chapters whose content is briefly summarised in the following list

Chapter 1 An overview on current challenges in kinetic modelling studies is given and the main goals of the research project are described.

Chapter 2 This Chapter organises and presents the state-of-the-art of kinetic model building and identification. It provides an introduction on the mathematical and statistical tools that will be used in the following research Chapters.

Chapter 3 A systematic approach to online model reparametrisation for robust parameter estimation in the presence of model sloppiness is presented in this Chapter. In the approach, the arising of model sloppiness is averted by optimally transforming the model parameter space every time new data are collected and included in the parameter estimation problem. The aim of online reparametrisation is to reduce the chance of numerical failures associated with model sloppiness without wasting experimental

resources and avoiding the introduction of bias in the parameter estimation problem. The approach is demonstrated both on a simulated and on a real case study where the aim is the estimation of parameters in a model of benzoic acid esterification with ethanol in a tubular reactor (Pipus et al., 2000). The modelling frameworks proposed in the following Chapters are formulated assuming that model parameters can be robustly estimated from available experimental data. The reparametrisation approach proposed in this Chapter may be coupled with any of the modelling algorithms illustrated in the following Chapters to improve their robustness at the parameter estimation and experimental design stages.

Chapter 4 A systematic approach for the online identification of approximated model structures is introduced in this Chapter. The central block in the procedure is a Model-Based Data Mining (MBDM) method for parameter estimation derived from robust regression theory (Rousseeuw and Leroy, 1987). MBDM generates two outputs: *i*) it classifies the explored experimental conditions as compatible or incompatible with the modelling hypotheses and *ii*) it estimates the model parameters excluding from the fitting the data that are incompatible with the modelling assumptions. A nonlinear support vector classifier (Cortes and Vapnik, 1995; Schölkopf and Smola, 2002) is then trained on the classified (observed) experimental conditions to build a reliability map, which quantifies the expected model reliability in unexplored experimental conditions. The generated maps can be employed to prevent the use of false optimal process points located in regions of low model reliability. Furthermore, an experimental design criterion to improve parameter precision in approximated models will be introduced in this Chapter where the design of experiments is *constrained* within the model reliability domain. The approach is demonstrated online in a simulated case study on the identification of an approximated model of ethanol dehydrogenation on copper-based catalyst (Carotenuto et al., 2013). A further case study is presented where the approach is applied offline for the identification of an approximated model of methanol oxidation on silver catalyst using real experimental data (Andreasen et al., 2005).

Chapter 5 If the available model is not reliable at the conditions of interest, a change in the model structure may be required. In this Chapter, a model building framework based on maximum likelihood inference is proposed where the structure of an avail-

able kinetic model is iteratively refined until an appropriate structure is obtained. In the proposed approach, model improvement is achieved in two steps: *1*) a step of model misspecification *diagnosis* and *2*) a step of model structure *evolution*. This Chapter focuses primarily on the former aspect. Statistical evidence provides an index to the scientist to justify changes in the model structure. Whenever over-fitting is detected, irrelevant free parameters are removed from the model structure. A Wald test (Wald, 1943) is employed to detect which parameters are unnecessary for fitting the data. If under-fitting is detected, the model structure is evolved by replacing relevant free parameters with state-dependent functions. A tailored Lagrange multipliers test (Silvey, 1959) is introduced in this manuscript to support the detection of promising parameters that should be considered for evolution. A Model Modification Index (MMI) is defined as a function of the Lagrange multipliers statistic and is proposed as a heuristic measure of model misspecification. The use of the MMI is illustrated in two simulated case studies on the diagnosis of model misspecification in a model of baker's yeast growth in a fed-batch bioreactor (Asprey and Macchietto, 2000) and in a model of glucose-insulin interaction in a healthy subject (Bergman et al., 1981).

Chapter 6 The main focus in this Chapter is the *evolution* of under-fitting model structures, i.e., the evolution of models in the presence of significant process-model mismatch. Under-fitting model structures are evolved by replacing some relevant model parameter with a state-dependent function. Relevant model parameters are detected by using a MMI-based approach. An Effect Relevance Index (ERI) is introduced as a function of the Lagrange multipliers statistic (Silvey, 1959) to support the scientist in the construction of opportune functional forms to replace model parameters. The use of the ERI is illustrated on the same simulated case studies used in Chapter 5.

Chapter 7 In this Chapter, the achievements of the research projects are summarised and possible future research directions are illustrated.

1.4 Computational resources

A number of computational resources were employed in this project and were instrumental for demonstrating the proposed computational frameworks and model identification techniques. Most of the numerical results presented in this Thesis were obtained using Python (Python Core Team, 2018). Python is a high-level programming language

that is widely used for numerical analysis. Some of its most established open-source libraries are extensively used throughout this work. The Python package *NumPy* `https://github.com/numpy/` (Oliphant, 2015) is employed for the manipulation of algebraic objects, i.e., arrays and matrices. Equations solvers and optimisation routines implemented in the library *SciPy* `https://github.com/scipy/` (Jones et al., 2001) are also used. The integration of systems of ordinary differential equations is performed using the equation solver LSODA (Petzold, 1983; Hindmarsh, 1992) implemented in *SciPy*. LSODA can solve initial value Cauchy problems with dense or sparse Jacobian. It also implements an automatic method to monitor the stiffness of the problem and choose which integration method to use (Petzold, 1983). The LSODA function implemented in the package *Scipy* (Mayorov et al., 2018) represents a wrapper to the LSODA routine that was originally implemented in the Fortran ODEPACK library, which can be accessed at `www.netlib.org/odepack/` (Hindmarsh, 2001).

Two numerical optimisation routines implemented in *SciPy* are employed in the work presented in this Thesis to solve parameter estimation and optimal experimental design problems, namely the gradient-free *Nelder-Mead* algorithm (Nelder and Mead, 1965) and the *SLSQP* solver (Nocedal and Wright, 2006). Assuming that the optimisation problem involves n optimisation variables, the Nelder-Mead method starts by building a simplex, namely a polytope of n+1 vertices. It evaluates the objective function at all vertices and then replaces the worst point with a new point, which is computed as a function of the polytope vertices at the current iteration (Nelder and Mead, 1965). In contrast to the Nelder-Mead method, the SLSQP algorithm can handle equality and inequality constraints. Since SLSQP requires the computation of second-order derivatives, both objective functions and constraints must be twice-differentiable with respect to the optimisation variables. With SLSQP, the step at a given iteration is computed by solving a quadratic optimisation program where the objective function includes the gradient and the second-order derivatives of the objective function and its constraints (Nocedal and Wright, 2006). The Python code for both the Nelder-Mead and the SLSQP algorithm can be found in `https://github.com/scipy/scipy/blob/master/scipy/optimize/`. The Python package *scikit-learn* (Pedregosa et al., 2011) includes a comprehensive library of Machine Learning models and it is employed in this work for the rapid implementation and training of Support Vector Machine Classifiers (Cortes and Vap-

nik, 1995; Schölkopf and Smola, 2002). The scikit-learn library can be downloaded from `https://github.com/scikit-learn/scikit-learn`.

In addition to Python, the software gPROMS® ModelBuilder developed by Process Systems Enterprise is used (PSE gPROMS, 2017). gPROMS® is a general-purpose modelling software for dynamic processes. It implements a comprehensive set of equation solvers, optimisation routines and model validation tools. In particular, the solver OAERAP (Outer Approximation Equality Relaxation Augmented Penalty) implemented in gPROMS® is used to solve robust regression problems with the aim of identifying outliers in kinetic datasets. The OAERAP solver is designed to solve both steady-state and dynamic optimisation problems with both continuous and discrete decision variables, i.e., Mixed-Integer NonLinear Programming (MINLP) optimisation problems. The OAERAP solver operates by performing a relaxation of the MINLP problem as a sequence of simpler optimisation problems, including NonLinear Programming (NLP) problems and Mixed-Integer Linear Programming (MILP) problems (Adjiman et al., 1998). Beside OAERAP, the differential and algebraic equations solver DASOLV implemented in gPROMS® is employed for the numerical integration of the dynamic models.

Unless differently stated, the numerical results presented in this work are obtained on a 64-bit Windows machine with processor Intel® Xeon® CPU E5-1650 v3 @ 3.50GHz and 32.0GB RAM. The Python scripts used to generate the numerical results presented in this Thesis can be accessed at the following repository `https://github.com/marcoquaglio92/quaglio_phd_thesis_code`.

Chapter 2

Literature survey

A survey of relevant literature on the topic of kinetic modelling is presented in this Chapter. A general form for the kinetic models considered in this Thesis is given in Section 2.1. An overview on different model classes is given in Section 2.2. Approaches available in the literature to build kinetic model structures are presented in Section 2.3. Statistical tools for bridging modelling and experimental activity are illustrated in Section 2.4. In the following sections, non-ideal scenarios in kinetic modelling studies are discussed, namely situations in which the model is affected by problems of practical identifiability (Section 2.5), cases in which collected data significantly deviate from the modelling assumptions (Section 2.6) and situations in which an available kinetic model structure has to be improved embracing the available experimental evidence (Section 2.7). In Section 2.8, a summary of the literature review is presented highlighting possible grey areas where additional research is required and where the work of this Thesis fits.

2.1 Deterministic models

The models considered in this work are in the form of mathematical laws stating a relationship among some variables of interest in the physical system and some parameters. Variables are quantities that may be either measurable or not and generally vary in both space and time. Parameters are assumed to be constant quantities that are not directly measurable. Throughout this Thesis, models are assumed to take the following general form

$$\begin{aligned} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t, \theta) &= \mathbf{0} \\ \hat{\mathbf{y}} &= \mathbf{h}(\mathbf{x}, \mathbf{u}, t, \theta) \end{aligned} \tag{2.1}$$

where $\mathbf{f}$ and $\mathbf{h}$ are respectively a $N_f \times 1$ and a $N_y \times 1$ array of model equations, $\mathbf{x}$ is an $N_x \times 1$ array of state variables, $\mathbf{u} \in U$ is a $N_u \times 1$ array of control input variables, t is time and array $\theta = [\theta_1, ..., \theta_{N_\theta}]^T \in \Theta$ represents a $N_\theta \times 1$ array of model parameters. $\hat{\mathbf{y}}$ represents a $N_y \times 1$ array of model predictions for a measurable set of system states $\mathbf{y}$. Possible constraints on the state variables are not explicitly stated in (2.1) for simplicity of notation. However, in general, sets of model equations may also include equality and inequality constraints on the state variables and/or on functions of the state variables. A model in the above form is called *deterministic* since the quantities appearing in its structure are assumed to be well determined in principle and not characterised by randomness (Bard, 1974). The modelling activity is concerned with selecting an opportune form for $\mathbf{f}$ and properly tuning the values of the parameters θ on the experimental observation. As opposed to *deterministic* models are *stochastic* models (Nelson, 1995). Quantities appearing in stochastic models are treated as inherently random and the simulation at the same conditions of a stochastic model produces different outcomes. Stochastic models are not considered throughout this work and more information about the application of stochastic modelling to chemical engineering related problems can be found in the literature (Diwekar and Rubin, 1991; Kristensen et al., 2004; Alshraideh and Runger, 2014).

2.2 Model classes

Every modelling assigment shall begin with the question: *Why is a model required?* In fact, depending on the final use of the model, some modelling strategies may be more opportune and effective than others. Furthermore, even for the same system, different models may be required depending on the specific application, e.g. process design (Biegler et al., 1997), process control (Ogunnaike and Ray, 1994) or optimisation (Pardalos and Resende, 2002). Following the distinction proposed by Bonvin et al. (2016), deterministic models are classified into three groups, depending on the factors that drive and determine their final structure.

Knowledge-driven models The structure of a knowledge-driven model is derived from a set of physically significant hypotheses. In process systems engineering, knowledge-driven models typically involve mass, momentum and energy balances and include phenomenological relationships to describe complex kinetic mechanisms (Rasmuson et al., 2014). These models are the most desirable because their structure organ-

ises the knowledge available on a system in a way that reflects the intrinsic causal mechanisms of the system itself. Knowledge-driven models possess some attractive characteristics:

- they provide insights on the fundamental degrees of freedom that are eventually responsible for the system behaviour (White et al., 2016).
- validated knowledge-driven models normally allow for extrapolation outside the range of data fitted for estimating their parameters (Bonvin et al., 2016).
- parameters in knowledge-driven models carry physical significance, which may help their estimation. As an example, the knowledge on a specific physical quantity (e.g. the kinetic rate of a known reaction, viscosity, *etc.*) may be transferred from a system to another.

The principal drawback of knowledge-driven models is that their development may require the investment of extensive amounts of time and resources for performing kinetic studies, formulate and select appropriate modelling hypotheses and translating them into a model structure.

Data-driven models There may be conditions in which it is impractical and/or unnecessary to capture the phenomenology of the system. In such situations, the aim is to identify an efficient mathematical description of the experimental observations, i.e., a regression curve or surface that well describes the distribution of the available data. This is the so called data-driven approach to modelling (Box and Draper, 1987). In data-driven modelling, parameters normally do not have physical significance and the model structure does not reflect the causal mechanisms of the physical system. As a direct consequence, data-driven models shall not be trusted when used to extrapolate the system behaviour beyond the conditions explored for their identification. A wide variety of data-driven model types were proposed in the scientific literature. These include response surfaces that were developed in the domain of statistical experimental design for studying the relationship between input factors and response variables (Box and Wilson, 1951), regression models based on artificial neural networks (Bishop, 1995), multivariate analysis methods such as principal component analysis (Jackson, 2003; Geladi and Kowalski, 1986) and kernel-based methods (Vapnik and Lerner, 1963; Smola and Schölkopf, 2004; Bah, 2008; Pillonetto et al., 2014).

Hybrid models A third class of models is identified by Bonvin et al. (2016) in the hybrid or grey-box models. In this model class, model structures are derived partially by physical and engineering knowledge, but including components that are constructed empirically from experimental observations. Typically, these components are built from data-driven models such as response surfaces (Bonvin et al., 2016), artificial neural networks (Cubillos et al., 2007; Xiong and Jutan, 2002) or kernel-based models (Del Rio Chanona et al., 2019). Data-driven components are included in the structure primarily for bridging the gap left by the incomplete knowledge available on the system (Brendel and Marquardt, 2008; Hof et al., 2009).

Disadvantages and merits of the three model classes are related to the effort required for the development of the model structure and to the extent to which the model captures the phenomenology of the system. Bonvin et al. (2016) also observed that methodologies for the identification of hybrid models were not pursued as systematically as for the other model classes. As a consequence, it frequently happens that when prior knowledge is not sufficient to build a knowledge-driven model, a purely data-driven model is built instead and the available insights on the system mechanisms are completely neglected.

2.3 Approaches for model structure building

Harnessing the complexity of kinetic phenomena into mathematical equations has been object of study for many scientists. Four main approaches for constructing phenomenological kinetic model structures were identified in the scientific literature: Bottom-up, Top-down, Superstructure-based and Incremental.

Bottom-up In the bottom-up approach, the starting point is a minimal reaction network where thermodynamically consistent rate laws are built using algebraic methods and/or graph theory (Marin and Yablonsky, 2011). Kinetic parameters may be then obtained from either a database of chemical properties (Song, 2004), theoretical calculations (Benson and Buss, 1958; Magoon and Green, 2013) or through the fitting of experimental data (Bard, 1974). If experimental evidence shows that the minimal mechanism does not represent the phenomenon with satisfactory accuracy, then the modelling activity shall proceed with the extension of the reaction network. Additional species and/or reactions are included in the model following thermodynamically consistent additivity rules (Benson and Buss, 1958). The majority of algorithms

for the automated generation of reaction networks implement a bottom-up approach (Ugi et al., 1993).

Top-down The application of a bottom-up approach to model building may lead to the construction of a very detailed and broad reaction network that is impractical and/or unnecessary for engineering purposes. For such reason, top-down approaches were developed to identify simplified models, out of complex mechanisms, without losing significant model descriptive capability. A top-down approach may aim at reducing the model size by removing irrelevant, slow reactions from the model and/or lumping fast reactions together. The most popular methods for model reduction are driven by sensitivity analysis (Seigneur et al., 1982), linear and nonlinear mixed-integer programming (Petzold and Zhu, 1999; Edwards et al., 2000; Bhattacharjee, 2003) and manifold learning (Chiavazzo et al., 2014).

Superstructure-based In superstructure-based kinetic modelling frameworks a model superstructure is first constructed including possible *lumped* reactions occurring in the system and possibly present chemical species. Reactions and species may be included in the superstructure even if their presence is only speculated. Optimisation methods are then used to identify the smallest set of reactions that is capable of representing the experimental observations according to a pre-set level of accuracy (Petzold and Zhu, 1999; Edwards et al., 2000; Wilson and Sahinidis, 2016; Tsay et al., 2017). Superstructure-based modelling methods were also employed in the context of data-driven kinetic modelling (Cozad et al., 2014).

Incremental In an incremental modelling framework the identification of the model structure proceeds in an incremental fashion towards increasing level of detail (Marquardt, 2005). The procedure begins with the determination of an opportune reaction network. Techniques based on target factor analysis are available in the literature for determining the set of independent lumped reactions occurring in the system in a way that is independent from the kinetic rates (Bonvin and Rippin, 1990; Amrhein et al., 1999). The following step involves the determination of the functional dependencies to describe the kinetic rate laws. The procedure requires high-resolution measurement techniques for extracting functional dependencies for the variables of interest without assuming structures for the kinetic laws. Once opportune model inversion

techniques are applied (Engl et al., 2000) it is possible to move towards a higher level of detail postulating structures for the kinetic rates and fit the parameters to the data. The advantage of the method is that the functional dependencies extracted from high-resolution datasets can be employed for running simulations even if fundamental knowledge on the reaction rate structures is lacking.

The application of any of these strategies to kinetic modelling shall always be coupled to a thoughtful experimental activity for estimating the model parameters and for validating the modelling hypotheses in the course of the model building process. The following section provides an overview on the statistical approaches proposed in the literature to bridge the gap between modelling and experimental activities.

2.4 Statistical model building and identification

A general framework for model identification linking modelling and experimental activities is presented and the fundamental steps in the procedure are detailed. A summary of the most significant approaches and frameworks available in the literature for supporting the construction of kinetic model structures is reported. Particular attention is given to knowledge-based and hybrid kinetic models, i.e. kinetic models embodying causal mechanistic knowledge on the process behaviour.

2.4.1 Bridging modelling and experimental activity

A possible framework for linking modelling and experimental activity is given in Figure 2.1. The framework proposed here does not cover all the possible pathways that a modelling activity can take. However, it is general enough for allowing the introduction of a set of fundamental modelling tools that shall be employed in modern modelling practice.

In general, kinetic modelling studies involve three main stages (Asprey and Macchietto, 2000; Galvanin, 2010):

1. *Preliminary analysis.* At this stage, the prior knowledge available on the system is translated into a set of candidate model structures. Approaches as the ones presented in Section 2.3 may be employed for the purpose. An identifiability analysis shall then be performed to assess whether it is possible a-priori to uniquely estimate the model parameters by fitting experimental data.

2. *Model discrimination.* An appropriate model structure is selected at this stage by challenging the available models against experimental evidence. If experimental ev-

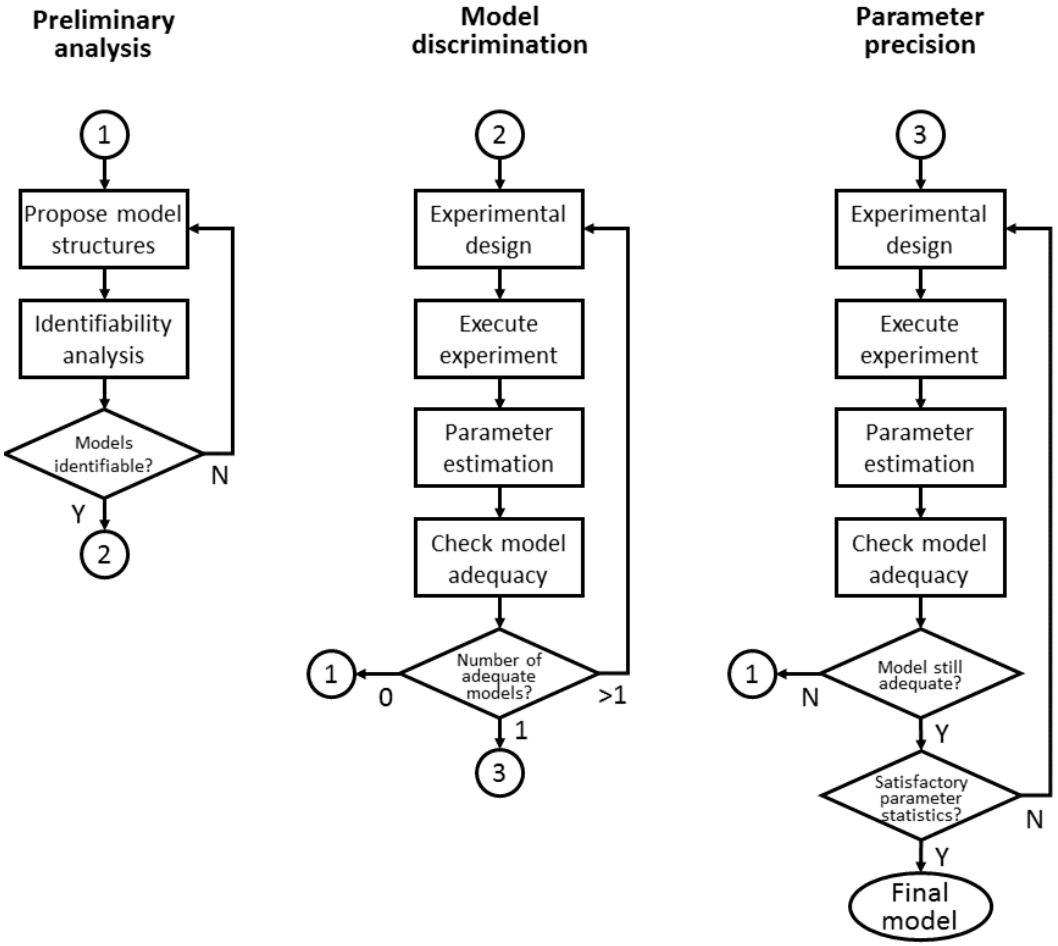


Figure 2.1: A general framework for model identification adapted from Asprey and Macchietto (2000).

idence suggests that none of the models is adequate, the modelling activity should start again from stage 1 with the formulation of a different set of model structures. If more than one model is adequate for describing the process, additional experiments may be performed with the aim of discriminating between rival model structures.

3. *Parameter precision.* When an adequate model structure has been selected, its identification requires the precise estimation of its kinetic parameters. If the available experimental data do not provide sufficient information to meet the desired statistical quality, additional experiments may be performed with the aim of maximising the collection of information for improving parameter precision.

The final model obtained at the end of the procedure shall be considered as a reasonably good representation of the physical system up to the moment when it is proved wrong by

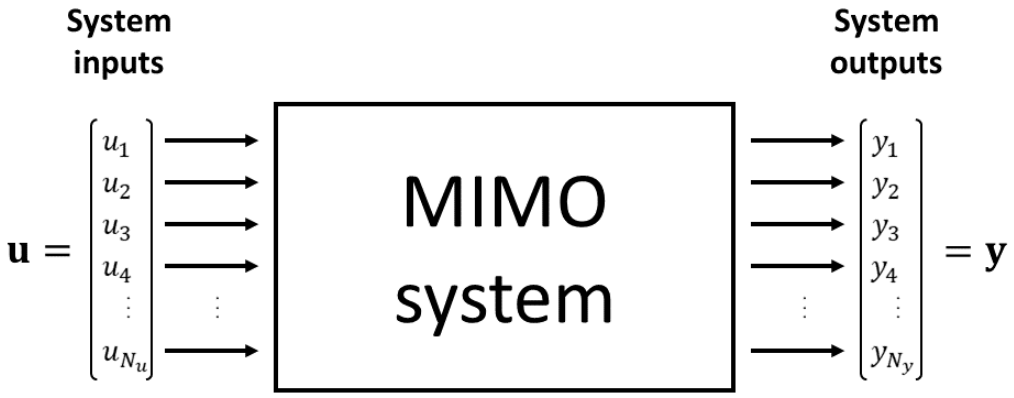


Figure 2.2: Generic Multiple Input Multiple Output (MIMO) system with N_u system inputs and N_y system outputs.

some new observations (Popper, 2002).

The tight interconnection between modelling and experimental activity needed in kinetic modelling studies requires the employment of a number of computational and statistical tools for identifiability analysis, experimental design, parameter estimation and model validation. A survey of the most relevant tools available for these purposes is given in the following sections. Particular emphasis is given to the concepts and tools that will be instrumental in the next research Chapters.

2.4.2 Identifiability analysis

In general, an experimental setup may be formalised mathematically as a Multiple Input Multiple Output (MIMO) system as shown in Figure 2.2 (Walter and Pronzato, 1997). In the figure, $\mathbf{u}$ is the array of system inputs and $\mathbf{y}$ is a $N_y \times 1$ array of measurable, output system states.

The inference problem is concerned with the selection of a model in the form (2.1) whose parameters can be uniquely identified from input-output measurements performed on the system. A formal definition of identifiability is now given.

Structural identifiability: A model in the form (2.1) is structurally *globally* identifiable from input-ouput data if for almost any parameter set $\theta^* \in \Theta$ there exists at least one input function $\mathbf{u}(t)$ such that the set of equations

$$\hat{\mathbf{y}}(\mathbf{u}, \theta) = \hat{\mathbf{y}}(\mathbf{u}, \theta^*) \tag{2.2}$$

admits the unique solution $\theta = \theta^*$ for all the possible initial values of the state variables $\mathbf{x}(0)$. A model is structurally *locally* identifiable if such condition is satisfied only in an open neighbourhood of the generic parameter set $\theta^* \in \Theta$ (Walter and Lecourtier, 1982; Saccomani et al., 2003).

In different words, if there exist at least two distinct parameter sets such that the input-output mapping described by the model is identical for any conceivable experiment then the model is deemed non-identifiable and cannot be employed for inference purposes. Structural identifiability is an intrinsic property of the system model structure (Walter and Lecourtier, 1982). Several systematic approaches were proposed in the literature to check if a model is identifiable. These can be classified as *a-priori*, which can be applied before any data is collected and *data-based* approaches (Raue et al., 2014). A further classification can be made between methods for *global* and *local* identifiability analysis.

A power series approach for testing local identifiability was proposed by Pohjanpalo (1978). The method aims at demonstrating that a model is structurally identifiable by analysing the power series of the time derivatives of the model outputs $\hat{\mathbf{y}}$ at $t = 0$. In the approach the following set of equations is constructed and constitutes the so called *exhaustive summary*

$$\frac{\mathrm{d}^k \hat{\mathbf{y}}(\theta)}{\mathrm{d}t^k}\Big|_{t=0} = \mathbf{a}_k(0) \quad \forall\, k = 1, ..., \infty \tag{2.3}$$

Using a differential algebra approach, a maximum order for the time derivatives can be obtained (Sedoglavic, 2002). The rank of the Jacobian associated with the exhaustive summary is then used as an index to assess which model parameters can be uniquely retrieved from input-output measurements (Karlsson et al., 2012). This method is known as the *Exact Arithmetic Rank* approach and it is implemented as a fully automatic function in Mathematica® (Wolfram Research, Inc., 2019). The approach can also be used to identify minimal sets of system outputs that guarantee a-priori model identifiability (Anguelova et al., 2012).

An approach for global identifiability was proposed by Ljung and Glad (1994) and is based on differential algebra. In this approach, model equations are manipulated in order to eliminate the non-observed states of the system. The input-output mapping is then parametrised linearly through a set of algebraic equations of the unknown parameter set θ. It is then possible to check using linear algebra algorithms whether this system of equations

admits a unique solution, which is a sufficient condition for global identifiability. The computational efficiency of the approach was improved by Audoly et al. (2001) and the method is now available for use in the system identification software DAISY (Bellu et al., 2007). Nevertheless, the application of such approaches remains impractical whenever the system of equations is large (more than 10 equations) (Raue et al., 2014).

An *optimisation-based* approach was proposed by Asprey and Macchietto (2000) as a mean to test identifiability in models involving a substantial number of equations. The approach aims at identifying two distinct parameter sets $\theta \in \Theta$ and $\theta^* \in \Theta$ such that the associated model predictions over an experimental time horizon τ are identical. Formally, the approach involves testing the following condition $\forall\, \mathbf{u} \in U$

$$\max_{\theta,\theta^* \in \Theta} [\theta - \theta^*]^T \mathbf{W}_\theta [\theta - \theta^*] < \varepsilon_\theta \tag{2.4}$$

$$\text{s.t.} \quad \int_0^\tau [\hat{\mathbf{y}}(\mathbf{u},\theta) - \hat{\mathbf{y}}(\mathbf{u},\theta^*)]^T \mathbf{W}_y [\hat{\mathbf{y}}(\mathbf{u},\theta) - \hat{\mathbf{y}}(\mathbf{u},\theta^*)] \mathrm{d}t < \varepsilon_y \tag{2.5}$$

where $\mathbf{W}_\theta$ $[N_\theta \times N_\theta]$ and $\mathbf{W}_y$ $[N_y \times N_y]$ are weighting matrices and ε_y and ε_θ are arbitrarily small positive real numbers. This approach has been successfully employed to test global identifiability in large scale biological models involving hundreds of parameters (Sidoli et al., 2005).

As observed by Saccomani et al. (2003), a-priori identifiability is a necessary condition for model-based inference. Nevertheless, even if a model satisfies the requirements for a-priori identifiability, it may still be impractical or impossible to estimate its parameters from noisy experimental data (Söderström and Stoica, 1989). To assess whether a model is identifiable in practice, *data-based* identifiability analysis methods may be employed. Data-based approaches are typically applied either with real experimental data or with simulated data if these can be generated under reasonable assumptions (Raue et al., 2014). The fundamental idea behind data-based approaches is that non-identifiability manifests as a flat fitting cost function in the parameter space (Raue et al., 2009). Data-based approaches to assess whether a model is identifiable from noisy experimental data will be discussed in Section 2.5 after the introduction of some additional concepts.

2.4.3 The parameter estimation problem

The estimation of the model parameters θ requires the fitting of experimental data. It is assumed that a dataset Y is available and it consists of N samples of $\mathbf{y}$. The dataset is denoted as follows

$$Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \tag{2.6}$$

Let $\varphi_i \in \Phi$ with $i = 1, ..., N$ denote the set of experimental conditions adopted for the collection of the i-th sample in Y. It is assumed that the measured quantities involved in a sample $\mathbf{y}$ are affected by Gaussian noise with known covariance Σ_y. A method that demonstrated to provide good estimates in a broad range of situations is the Maximum Likelihood (ML) estimator (Bard, 1974). As observed by Akaike (1998), the ML estimator aims at identifying the value of the parameters, namely the maximum likelihood estimate $\hat{\theta} = [\hat{\theta}_1, ..., \hat{\theta}_{N_\theta}]^T$, which minimises the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the nominal and the postulated distribution of the data. In different words, the ML estimate is the value of parameters that minimises the discrepancy between the distribution predicted by the model and the actual distribution of the data (White, 1982). The computation of the maximum likelihood estimate is performed through the maximisation of the likelihood function or, indifferently, its natural logarithm, which will be denoted with the symbol $\mathscr{L}$. The optimisation of the log-likelihood function $\mathscr{L}$ frequently reduces the numerical complexity of the problem (Bard, 1974).

$$\begin{aligned} \mathscr{L}(Y|\theta) = &-\frac{N}{2}[N_y \ln(2\pi) + \ln(\det(\Sigma_y))] \\ &-\frac{1}{2}\sum_{i=1}^{N}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)] \end{aligned} \tag{2.7}$$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathscr{L}(Y|\theta) \tag{2.8}$$

In (2.7), $\hat{\mathbf{y}}_i$ represents the model prediction for the i-th sample $\mathbf{y}_i$. The ML estimate satisfies the unconstrained maximum likelihood equations

$$\nabla \mathscr{L}(Y|\hat{\theta}) = \mathbf{0} \tag{2.9}$$

where the symbol ∇ denotes the gradient operator in the parameter space. The ML estimator

is consistent, i.e. if the model (2.1) is identifiable and correctly specified, the maximum likelihood estimate $\hat{\theta}$ exhibits a convergent behaviour as the number of fitted samples tends to infinity (Bard, 1974).

Many other popular estimators such as the Least Squares or the Weighted Least Squares estimators represent special cases of the ML estimator under the assumption of normally distributed measurement noise. Alternative estimators based on Bayesian inference were also proposed in the literature, e.g. the Maximum A Posteriori (MAP) estimator $\hat{\theta}_{MAP}$ (Bassett and Deride, 2019). The MAP estimator may be interpreted as a ML estimator which also accounts for prior knowledge on the possible values of the model parameters. Such prior knowledge is provided as an input to the problem in the form of a prior distribution $p(\theta)$ defined on the parameter domain Θ (Sorenson, 1980).

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} \mathscr{L}(Y|\theta) + \ln(p(\theta)) \tag{2.10}$$

2.4.4 Statistical tools for model validation

Once the model parameters have been optimally tuned to the experimental data, model adequacy is checked by performing statistical tests. Tests may be performed for different purposes: *1*) diagnosing the presence of modelling errors through the detection of *over-fitting* or *under-fitting*; *2*) assessing whether parameter estimates have been estimated with sufficient precision; *3*) selecting the best model available from the set of possible candidate models. This section is dedicated to a description of statistical methods for performing the aforementioned tasks.

2.4.4.1 Goodness of fit test

A model may be evaluated on its ability of producing predictions that are *close* to the corresponding experimental data. However, while assessing the quality of a model fitting, one shall take into account that measurements are inherently affected by measurement noise. The assessment of the model fitting involves a comparison between the distribution of the model residuals with the hypothetical distribution of the measurement errors. This may be performed through a χ^2 test (Silvey, 1975). The test aims at quantifying the probability of observing a certain distribution of residuals under the null hypothesis, i.e. under the hypothesis that the proposed model is correctly specified (Devore, 2010). The test starts from the hypothesis that the model residuals for $\theta = \hat{\theta}$ asymptotically follow the same distribution of the measurement noise, namely a multivariate Gaussian distribution with mean $\mathbf{0}$ and

covariance Σ_y.

$$\mathbf{y}_i - \hat{\mathbf{y}}_i(\hat{\theta}) \sim \mathcal{N}(\mathbf{0}, \Sigma_y) \quad \forall i = 1, ..., N \tag{2.11}$$

From the assumption (2.11) it is derived that the sum of the squared normalised residuals χ^2_Y is a random variable that is distributed as a χ^2 statistic where the appropriate degree of freedom in the presence of finite datasets is $N \cdot N_y - N_\theta$.

$$\chi^2_Y = \sum_{i=1}^{N} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\hat{\theta})]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\hat{\theta})] \sim \chi^2_{N \cdot N_y - N_\theta} \tag{2.12}$$

A two-tailed goodness-of-fit test based on χ^2 test may be employed to detect whether model residuals are too small or too large to be explained only with measurement noise. The two-tailed χ^2 test with significance α may have three possible outcomes which are summarised in the following

$$\text{Goodness-of-fit test} : \begin{cases} \chi^2_Y < \chi^2\left(\frac{1-\alpha}{2}\right) & \text{Failed for over-fitting} \\ \chi^2\left(\frac{1-\alpha}{2}\right) < \chi^2_Y < \chi^2\left(\frac{1+\alpha}{2}\right) & \text{Passed} \\ \chi^2\left(\frac{1+\alpha}{2}\right) > \chi^2_Y & \text{Failed for under-fitting} \end{cases} \tag{2.13}$$

where $\chi^2(\cdot)$ represents the percentile of a χ^2 distribution with degree of freedom $N \cdot N_y - N_\theta$ and the argument in brackets represents the level of significance. When the test is failed for over-fitting, the probability of observing residuals equal or smaller than χ^2_Y is low and typically indicates that the model involves an excessive number of free parameters, i.e. an excessive number of degrees of freedom to capture the trend underlying the available noisy data. When the test is failed for under-fitting, the probability of observing residuals equal or larger than χ^2_Y is low and indicates that the model structure may be inappropriate to model the system. In PSE, the presence of under-fitting is also referred to as the presence of process-model mismatch (Lee et al., 1989; Fotopoulos et al., 1996; Meneghetti et al., 2014). If the test is passed, the model structure may be considered as an appropriate description of the system.

2.4.4.2 Statistical quality of parameter estimates

The characterisation of the parameter estimates requires the computation of a confidence region in the parameter space. Under the assumption of Gaussian measurement noise, the

covariance matrix $\mathbf{V}_\theta$ of the parameter estimates is well approximated by the inverse of the observed Fisher Information Matrix (FIM) $\mathbf{H}$ (Walter and Pronzato, 1997).

$$\mathbf{V}_\theta \simeq \mathbf{H}^{-1} \tag{2.14}$$

The quality of the above approximation improves as the variance of the measurement noise decreases and the fitting of the model improves (Bard, 1974). The observed FIM $\mathbf{H}$ may be computed as the negative Hessian of the log-likelihood function $\mathscr{L}$ evaluated at the maximum likelihood estimate $\hat{\theta}$ (2.15).

$$\mathbf{H} = -\nabla\nabla^T \mathscr{L}(Y|\hat{\theta}) \tag{2.15}$$

From the covariance $\mathbf{V}_\theta$, it is possible to derive the confidence intervals for the estimates $\hat{\theta}$ and the correlation coefficient c_{ij} between any estimated parameter pair $\hat{\theta}_i$ and $\hat{\theta}_j$ (Bard, 1974). Let $v_{\theta,ij}$ be the ij-th element of the covariance matrix $\mathbf{V}_\theta$. The confidence interval with significance α for the i-th parameter estimate $\hat{\theta}_i$ can be computed as $\hat{\theta}_i \pm z_{\alpha/2}\sqrt{v_{\theta,ii}}$ where $z_{\alpha/2}$ represents a two-tailed value computed from a standard normal distribution with significance α. The correlation coefficient between any parameter pair $\hat{\theta}_i$ and $\hat{\theta}_j$ can be computed according to

$$c_{ij} = \frac{v_{\theta,ij}}{\sqrt{v_{\theta,ii}v_{\theta,jj}}} \quad \forall\, i,j \tag{2.16}$$

The statistical quality of the parameter estimates $\hat{\theta}$ can be checked through a one-tailed t-test with opportune significance α (Walpole et al., 2011). The test involves the computation of t-values for all parameters and their comparison with the t-value of reference as follows

$$\frac{\hat{\theta}_i}{t\left(\frac{1+\alpha}{2}\right)\sqrt{v_{\theta,ii}}} \geq t(\alpha) \quad \forall\, i = 1,...,N_\theta \tag{2.17}$$

where $t(\cdot)$ represent the t-values obtained from a Student's distribution with degree of freedom equal to $N \cdot N_y - N_\theta$ and significance given by the argument in brackets. If conditions (2.17) are satisfied for all parameters this may be interpreted as an index of satisfactory parameter precision.

The evaluated covariance matrix $\mathbf{V}_\theta$ can be used to compute an approximated confidence region for the parameter estimates in the form of a confidence ellipsoid. The 95%

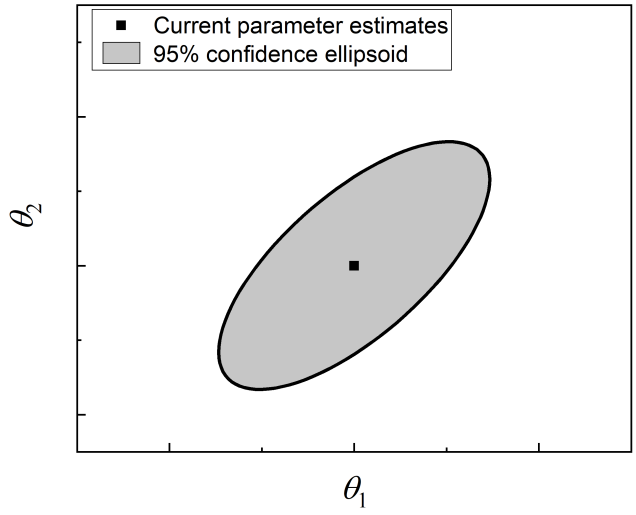


Figure 2.3: Parameter estimates for the parameter pair $\theta_1 - \theta_2$ and associated 95% confidence ellipsoid.

confidence ellipsoid is computed as the set of parameters θ that satisfy the following condition

$$[\theta - \hat{\theta}]^T \mathbf{V}_\theta^{-1} [\theta - \hat{\theta}] \leq \chi^2_{N_\theta}(95\%) \tag{2.18}$$

An example of how confidence ellipsoids may be represented graphically is given in Figure 2.3, where the confidence region is projected on the $\theta_1 - \theta_2$ plane in the parameter space. Alternative approaches for constructing more accurate parameter inference regions accounting for possibly complex geometries in the log-likelihood profile are also available in the literature (Benabbas et al., 2005).

2.4.4.3 Model structure selection

Whenever multiple models are proposed to describe a given phenomenon, the scientist may be interested in comparing their performance in representing the experimental observations. A number of information-theoretic criteria have been proposed to select the best available model assuming that the goal is the identification of an optimal compromise between fitting quality and model complexity, which is quantified by the number of free parameters N_θ. The Akaike Information Criterion (AIC) was proposed by Akaike (1974) as a relative measure of the information loss associated with the selection of a certain model compared to another. The AIC index associated with a given model structure is evaluated as follows

$$\text{AIC} = 2N_\theta - 2\mathscr{L}(Y|\hat{\theta}) \tag{2.19}$$

The AIC index may be evaluated for all the available models and the model with the lowest AIC index may be selected. Statistical tests based on the AIC are available to assess whether there is sufficient evidence to state that a model is significantly *better* than another (Sakamoto et al., 1986). The AIC may be interpreted as a relative measure of merit that accounts both for the model fitting quality but penalises models with a high number of parameters.

An alternative criterion based on Bayesian inference was derived by Schwarz (1978), i.e. the Bayesian Information Criterion (BIC). The criterion proposed by Schwarz (1978) was derived from the Bayesian Occam's razor under a number of simplifying hypotheses (Barber, 2011). The BIC index of a given model is evaluated as follows

$$\mathrm{BIC} = N_\theta \ln(N \cdot N_y) - 2\mathscr{L}(Y|\hat{\theta}) \tag{2.20}$$

The BIC index shall be evaluated for all the available models and the model with the smallest BIC may be selected as the *best* available structure. Analogously to the AIC, statistical tests can be performed to assess whether the BIC of a given model is significantly smaller than the BIC of a competing model structure (Burnham and Anderson, 2002). As observed by Burnham and Anderson (2004), the BIC tends to penalise more than the AIC the presence of a higher number of parameters in the model. For such reason, it is often stated that the BIC criterion is more conservative than the AIC (Burnham and Anderson, 2004).

2.4.5 Model-based design of experiments for model discrimination

The application of the statistical tools presented in Section 2.4.4 may not allow a clear-cut selection of a single appropriate model. As an example, it may happen that two model structures derived from two sets of irreconcilable modelling hypotheses cannot be both true at the same time. Nonetheless, it may happen that neither model is falsified by the goodness-of-fit test given the available experimental evidence. If multiple models are adequate in representing the dataset Y the scientist shall proceed by performing additional experiments with the aim of discriminating between the competing model structures (Hill, 1978).

Perhaps, the first who posed the problem of designing experiments for model discrimination was Cox (1961, 1962). However, his first scientific contributions were more focused on the problem of model selection for a given dataset rather than on the design of experiments for model selection. In a later work, Chambers and Cox (1967) extended the work of Cox (1961) including an experimental design step, but the problem was addressed only in

the context of distinguishing between a logistic and binary response models.

Hunter and Reiner (1965) were the first who considered the problem of designing samples with the aim of discriminating between rival kinetic models. They proposed a criterion for discriminating between two rival models that is based on the design of samples at conditions where the divergence between model predictions is largest (Hunter and Reiner, 1965). The work of Hunter and Reiner (1965) was extended to the case where there are three or more competing model structures by Roth (1966). The design criterion proposed by Hunter and Reiner (1965) was introduced only on an intuitive basis and was later formalised by Atkinson and Fedorov (1975a,b).

Box and Hill (1967) and Fedorov and Pázman (1968) observed that the criterion proposed by Hunter and Reiner (1965) failed to consider the variance on the model responses. They extended the work including the uncertainty on the model predictions in criteria for MBDoE for model discrimination. In particular, Box and Hill (1967) proposed a Bayesian approach were the design is performed by evaluating the posterior probabilities that the available models are appropriate to model the phenomenon. In a later work, Hsiang and Reilly (1971) improved the approach considering the possibility that model predictions and measurement noise may not follow normal distributions.

A criterion based on frequentist statistic was proposed by Buzzi-Ferraris and Forzatti (1983). The proposed approach does not require the computation of posterior probabilities and incorporates criteria to assess whether the level of measurement noise is excessive for model discrimination purposes. The approach was later refined and extended to the case of multiresponse models (Buzzi-Ferraris et al., 1984, 1990). In a later work, Schwaab et al. (2006) extended the work of Buzzi-Ferraris et al. (1984) including Bayes factors in the objective function to quantify the probability that a given model is the most appropriate representation of the system. Schwaab et al. (2008b) also proposed an improved version of this criterion where the influence of the designed experiment on the posterior covariance of the model predictions is considered in the design stage.

Once the discriminant experiment is performed, additional collected samples are included in the dataset. The parameters of the competing models are re-estimated and model adequacy is checked again with a goodness-of-fit test and/or with statistical tests based on information criteria (see Section 2.4.4). The collection of samples shall continue until a single model is found adequate for describing the physical phenomenon.

2.4.6 Model-based design of experiments for parameter precision

Once a single model structure is selected as an adequate representation of the system, further validation procedures may aim at reducing the uncertainty on the parameter estimates, i.e. improving parameter precision. The precise estimation of the model parameters relies on the fitting of measurements collected at conditions where model predictions are most sensitive to a change in the parameter values (Box and Lucas, 1959). This sensitivity can be interpreted as the information that measurable system states carry for the estimation of non-measurable model parameters and it is quantified by the Fisher Information Matrix (FIM) (Walter and Pronzato, 1997). A variety of Model-Based Design of Experiments (MBDoE) criteria have been proposed in the literature to design information-rich experiments accounting for the limited amount of resources available for the experimentation (Espie and Macchietto, 1989; Prasad and Vlachos, 2008; Dirion et al., 2008). These approaches are based on the computation of the expected covariance matrix of the model parameters $\hat{\mathbf{V}}_\theta$ as a function of the experimental conditions of the samples to be designed. Let N_{sp} denote the number of samples that the scientist is willing to design for improving parameter precision. Furthermore, let φ_k with $k = 1,...,N_{sp}$ be the experimental conditions associated with the samples to be designed. The predicted covariance matrix is evaluated as follows

$$\hat{\mathbf{V}}_\theta(\varphi_1,...,\varphi_{N_{sp}}) = \left[\mathbf{V}_\theta^{-1} + \sum_{k=1}^{N_{sp}} \hat{\mathbf{H}}_k\right]^{-1} \tag{2.21}$$

where the first addend in the brackets quantifies the information that is available from previously fitted samples and the second addend quantifies the expected information associated with the samples that are yet to be collected. The quantity $\hat{\mathbf{H}}_k$ appearing in (2.21) represents the expected FIM associated with the k-th sample under design and it is evaluated according to the following expression

$$\hat{\mathbf{H}}_k = \nabla\hat{\mathbf{y}}(\varphi_k,\hat{\theta})\Sigma_y^{-1}\nabla\hat{\mathbf{y}}(\varphi_k,\hat{\theta})^T \quad \forall k = 1,...,N_{sp} \tag{2.22}$$

The MBDoE problem is then recast in terms of minimising the expected confidence region of the parameters after the conduction of the experiment to be designed. In order to summarise the multidimensional nature of $\hat{\mathbf{V}}_\theta$ into a scalar quantity, different measures ψ of $\hat{\mathbf{V}}_\theta$ were proposed in the literature as objective functions to be minimised for an optimal MBDoE. The most popular design criteria are (Pukelsheim, 2006):

- *A-optimal*: the objective function is $\psi = \mathrm{Tr}(\hat{\mathbf{V}}_\theta)$ and it is equivalent to minimising the volume of the rectangular hyper-box that contains the expected confidence ellipsoid of the parameter estimates;
- *D-optimal*: where the objective function chosen for minimisation is $\psi = \mathrm{Det}(\hat{\mathbf{V}}_\theta)$ and it corresponds to minimising the volume of the expected confidence ellipsoid in the parameter space;
- *E-optimal*: this criterion aims at minimising the largest eigenvalue of $\hat{\mathbf{V}}_\theta$ and it is equivalent to minimising the longest axis of the expected confidence ellipsoid;

A geometric interpretation of these design criteria is given in Figure 2.4. The above list of design metrics is by no means complete. For a more comprehensive review of design criteria for parameter precision the reader is referred to the relevant literature (Pukelsheim, 2006; Franceschini and Macchietto, 2008b; Galvanin, 2010).

Once an appropriate scalar measure ψ is selected, optimal MBDoE problems are recast as an optimisation problem in the following form

$$\begin{aligned} \varphi_1^*, ..., \varphi_{N_{sp}}^* &= \arg \min_{\varphi_1,...,\varphi_{N_{sp}}} \psi(\varphi_1, ..., \varphi_{N_{sp}}) \\ \text{s.t.} \quad \varphi_k &\in \Phi \quad \forall k = 1, ..., N_{sp} \end{aligned} \tag{2.23}$$

and are typically solved by using numerical optimisation routines. Once the above optimisation problem is solved, from the predicted covariance $\hat{\mathbf{V}}_\theta(\varphi_1^*, ..., \varphi_{N_{sp}}^*)$ it is possible to derive an approximation for the predicted confidence intervals for the model parameters.

One shall observe that whenever the model is nonlinear in the parameters, the expected information matrix $\hat{\mathbf{H}}$ is a function of the parameter values. The expected FIM is therefore evaluated at the best estimate available, namely the ML estimate $\hat{\theta}$. Nonetheless, parameter estimates may be highly uncertain at the experimental design stage. MBDoE approaches that are robust towards parameter uncertainty have been proposed in the literature (Asprey and Macchietto, 2002; Körkel et al., 2004; Bruwer and MacGregor, 2006; Mesbah and Streif, 2015).

2.5 Practical identifiability and model sloppiness

Even if a model structure satisfies the requirements for a-priori identifiability (see Section 2.4.2), it may still be impossible or extremely challenging to precisely estimate its parameters by fitting noisy experimental data (Transtrum et al., 2010). Let $\lambda_1, ..., \lambda_{N_\theta}$ denote the

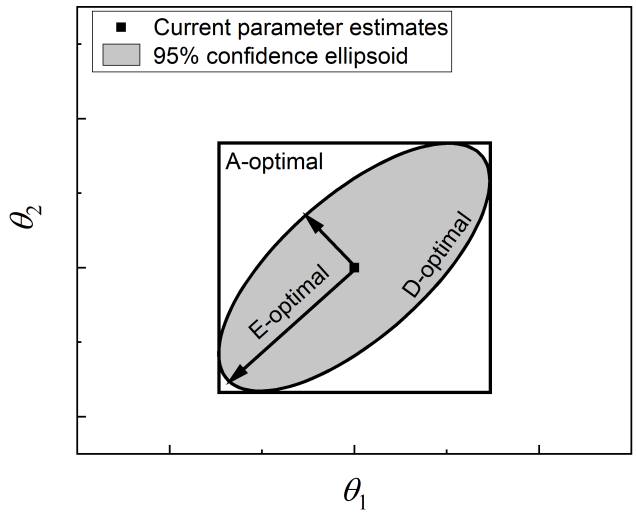


Figure 2.4: Geometric interpretation of the experimental design criteria. Figure adapted from Franceschini and Macchietto (2008b), Galvanin (2010) and Fedorov and Leonov (2013).

eigenvalues of the observed FIM **H**. The ratio between the maximum and the minimum eigenvalue represents the condition number of the model identification problem κ

$$\kappa = \frac{\max_i \lambda_i}{\min_i \lambda_i} \tag{2.24}$$

It may happen that the smallest eigenvalue of **H** is below unity. When such circumstance occurs, the model is considered non-identifiable given the available dataset (White et al., 2016). In fact, White et al. (2016) observed that changes in the parameter values along the directions associated with eigenvalues smaller than unity tend to produce changes in the model predictions that are irrelevant compared to the level of measurement noise in the system. It may also happen that the eigenvalues of **H** span over a wide range of orders of magnitude. Whenever these circumstances occur the condition number may be extremely high and the model is called *sloppy* (Chiş et al., 2014). As shown in Figure 2.5, model sloppiness manifests in confidence ellipsoids characterised by extremely high eccentricity (Raue et al., 2009). It is recognised that kinetic models of chemical and biochemical phenomena frequently exhibit a sloppy behaviour (White et al., 2016). Parameter estimation and optimal MBDoE problems are normally recast as optimisation problems and solved numerically. In the presence of a nearly singular information matrix and/or a high condition number, numerical optimisation routines may fail in solving the aforementioned problems. More information on the possible numerical failures associated with model sloppiness can be found in Section 1.2.2.1.

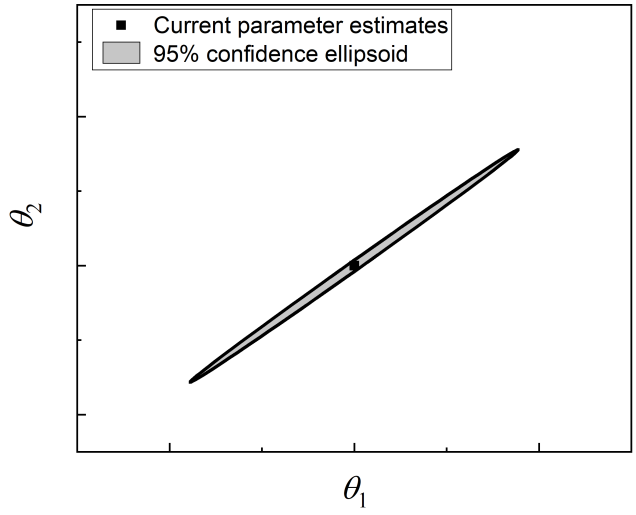


Figure 2.5: In the presence of model sloppiness, the eigenvalues of the FIM **H** and consequently the eigenvalues of the parameter covariance $\mathbf{V}_\theta$ span over a wide range of scale lengths. This typically results in confidence ellipsoids characterised by high eccentricity and a high chance of numerical failures when numerical model identification routines are invoked.

Several approaches have been proposed in the literature to address the practical identifiability problems associated with sloppy models and mitigate the risk of numerical failures associated with practical identifiability issues (Dovi et al., 1994):

1. *Experimental-design-based (ED) methods.* These methods are based on the design of optimal experiments for *reshaping* the covariance matrix of the parameter estimates and improve the condition number. For more information on these approaches, the reader is referred to the relevant literature on design criteria for relaxing model sloppiness and reducing parameter correlation (Hosten, 1974; Pritchard and Bacon, 1978; Versyck and Van Impe, 1997; Galvanin et al., 2007; Franceschini and Macchietto, 2008a,c,d; Maheshwari et al., 2013; Chiş et al., 2014; Wilson et al., 2015; Shahmohammadi and McAuley, 2019).

2. *Regularisation-based (RG) methods.* Regularisation involves the introduction of a bias in the parameter estimates with the aim of constraining their variance and, concomitantly, reducing the condition number associated to the parameter estimation problem (Barz et al., 2016). Popular regularisation techniques are *i*) Tikhonov regularisation (Johansen, 1997; Hansen, 2005; Bardow, 2008) *ii*) truncated singular value decomposition (Hansen, 2005; López Cárdenas et al., 2015) and *iii*) parameter subset selection (Barz et al., 2013; López Cárdenas et al., 2015).

3. *Reparametrisation-based (RP) methods.* The aim of reparametrisation is transforming the original parameter space Θ into a robust parameter space Ω where both parameter estimation and MBDoE can be performed more effectively on well-conditioned objective functions (Agarwal and Brisk, 1985a,b). Although there is no theoretical advantage in the use of a reparametrised model (Rimensberger and Rippin, 1986; Dovi et al., 1994), the performance of model identification algorithms is sensitive to the type of parametrisation used (Espie and Macchietto, 1988). The effectiveness of RP-based methods has been recognised in many kinetic studies in the literature (Espie and Macchietto, 1988; Asprey and Naka, 1999; Benabbas et al., 2005; Schwaab and Pinto, 2007; Schwaab et al., 2008a; Buzzi-Ferraris and Manenti, 2009).

These methods present strengths and weaknesses. ED-based methods are systematic. Optimal ED criteria to relax model sloppiness can be easily implemented into a computer program. However, even optimally designed experiments may not be sufficient to bring the condition number below critical levels. This weakness of ED-based methods is typically associated with either a too narrow range of explorable experimental conditions and/or an insufficient experimental budget to perform these optimal experiments. Furthermore, optimally designed experiments to reduce the condition number may not carry optimal amounts of information for the estimation of the model parameters. This limitation may be overcome by designing experiments that represent a compromise between improving the parameter statistics and reducing the condition number (Franceschini and Macchietto, 2008c; Maheshwari et al., 2013).

An advantage of RG-based and RP-based methods is that they do not require the execution of experiments for improving the condition number and one can devote the entire experimental budget on improving the statistics of the parameter estimates. In RG-based approaches, the condition number is controlled through the introduction of prior information on the model parameter values in the form of a prior parameter distribution. Systematic approaches, e.g. approaches based on Bayesian inference (MacKay, 1992), are available in the literature for supporting the selection of appropriate priors (Hansen, 2005). The introduction of prior information in the parameter estimation problem generally results in the computation of biased parameter estimates.

In contrast to RG-based approaches, RP-based methods do not involve the introduction of any bias in the model identification problem. Ad hoc strategies to reparametrise sloppy

models were suggested for very specific kinetic model structures, e.g. Arrhenius-type reaction rates (Asprey and Naka, 1999; Schwaab and Pinto, 2007; Schwaab et al., 2008a; Buzzi-Ferraris and Manenti, 2009). However, only few systematic approaches to the reparametrisation of sloppy models are available in the literature (Espie and Macchietto, 1988). An additional feature of RP-based methods is that whenever a model is reparametrised, the parametrisation is fixed until the end of the experimental campaign. However, sloppiness is a consequence of the combination of both the model parametrisation and the dataset available to identify the model (Söderström and Stoica, 1989). There is no theoretical guarantee that the reparametrised model will not become sloppy after the collection of new data (Wilson et al., 2015). The arising of sloppiness may be averted by adjusting the parametrisation online in the course of the experimental activity, i.e. by reparametrising the model every time new data are collected and included in the parameter estimation problem. Nonetheless, online applications of RP-based methods seem to be missing in the scientific literature.

2.6 Robust regression and outlier detection

In kinetic modelling studies it frequently happens that some observations significantly deviate from the modelling assumptions (Huber, 1981). These observations represent outliers and their presence in the dataset may be explained with three possible causes:

1. *Significant system disturbances.* External disturbances occurring in the course of the experiments and/or in the course of the sampling process may lead to the collection of data affected by a measurement noise that deviates from the postulated assumption of Gaussian noise with zero mean and covariance Σ_y.

2. *Systematic errors.* The presence of offsets in the setup control system and/or an inaccuracy in following the experimental protocol may lead to the collection of data that deviate significantly from the nominal behaviour of the system.

3. *Inappropriate modelling assumptions.* The model structure used to fit the data may be inappropriate to represent the behaviour of the system across the entire domain of experimental conditions.

A fraction of outliers in the dataset between 1% and 10% shall always be expected (Huber, 1981). When classic estimators are employed, the presence of outliers in the dataset inevitably leads to biased parameter estimates (Rousseeuw and Leroy, 1987). The concept

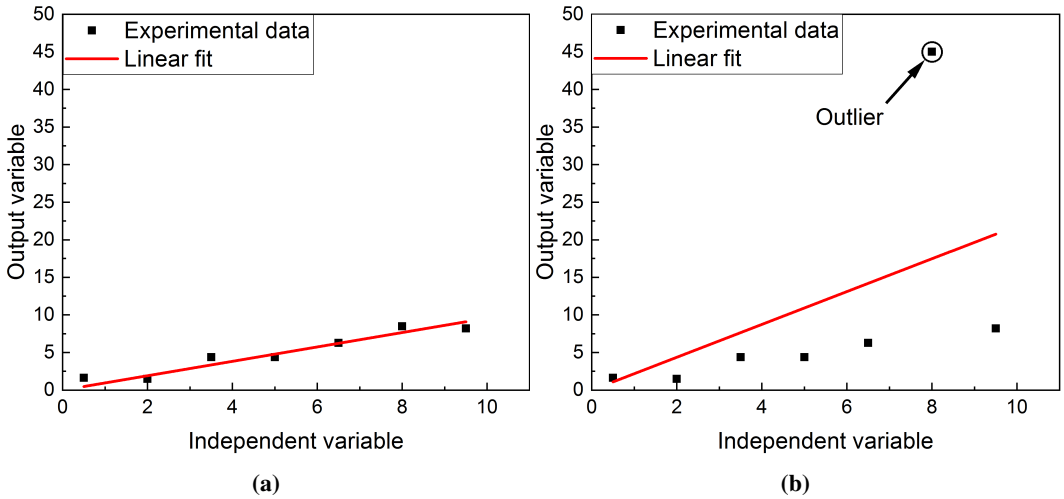


Figure 2.6: Effect of a single outlier on the Least Squares Estimator. (a) The dataset does not contain outliers and the estimated slope in the linear fit is unbiased. (b) There is a single outlier in the dataset and the slope estimated through the linear fit is biased. If the value of the outlier were increased to infinity also the estimated slope would increase beyond any finite bound.

of asymptotic breakdown point is used to quantify the robustness of an estimator to outlier contamination in the dataset. The breakdown point is defined as the minimum fraction of outlier contamination in the dataset that can carry the parameter bias beyond any finite bound (Hampel, 1985).

For standard ML estimators, the asymptotic breakdown point is 0%, meaning that the presence of a single outlier in the dataset can result in an unbounded bias on the parameter estimates. This is illustrated in Figure 2.6 for the Least Squares Estimator. Figure 2.6a shows a linear fit in the absence of outliers, where the estimated slope associated with the linear regression is unbiased. In Figure 2.6b, one of the points is replaced with an outlier and the slope computed through linear regression is biased. If the value of the outlier were brought up to infinity, the estimated slope would also be carried beyond any finite bound. The acknowledgement of this weakness led to the development of different estimators capable of handling outlier contamination in the dataset and a whole sub-field of statistics, namely the field of robust regression. Rousseeuw and Leroy (1987) attribute the first important step in the field of robust regression to Edgeworth (1887). He observed that estimators based on least squares minimisation are particularly sensitive to outliers because residuals are squared in the objective function. He therefore proposed to minimise the sum of the absolute values of the residuals. Such estimator is known as the L_1 estimator. Nonetheless, Rousseeuw and Leroy (1987) showed that the L_1 estimator reduces the parameter bias but

its breakdown point is still 0%.

Significant contributions in the field of robust regression are attributed to Huber (1981) with the development of the so-called M-estimators. In this class of estimators, the objective function is the sum of an appropriate function ρ of the residuals. Such function ρ may be appropriately selected to exclude from the fitting data that are in disagreement with the modelling assumption. In its seminal work on M-estimators, Huber (1973) considered only single-response linear models and the problem of estimating location parameters in multivariate Gaussian populations. A M-estimator $\hat{\theta}_M$ for multi-response linear models was studied by Collins (1982) in the following form

$$\hat{\theta}_M = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} \rho\left([\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]\right) \tag{2.25}$$

As observed by Rousseeuw and Leroy (1987), M-estimators still suffer from a breakdown point of 0% due to their sensitivity to outliers caused by offsets in the explanatory variables. This led to an extension of the approach proposed by Huber that resulted in the development of generalised M-estimators, also known as GM-estimators or bounded-influence estimators (Mallows, 1975; Hill, 1977). The main idea behind GM-estimators was to bound the effect of outliers in the explanatory variables by including weights in the objective function. However, Maronna et al. (1979) demonstrated that even for GM-estimators, the breakdown point cannot be above a certain threshold and decreases as the number of parameters increases.

A different approach, namely the *repeated median* method was proposed by Siegel (1982). The repeated median estimator can achieve a breakdown point of 50%, which is the maximum fraction of outlier contamination that an estimator can theoretically handle (Hampel, 1985). However, the approach proposed by Siegel (1982) involves a check on all the possible subset of samples in the dataset and the method may rapidly become computationally intractable. From the idea of Siegel (1982), Rousseeuw (1984) developed the *Least Median of Square* LMS estimator where the sum of squared residuals is replaced by the median operator

$$\hat{\theta}_{LMS} = \arg\min_{\theta\in\Theta} \text{median}_i\left([\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]\right) \tag{2.26}$$

The LMS estimator still possesses the property of the high 50% breakdown point and

was also tested on nonlinear regression problems (Stromberg, 1993). Nonetheless, the LMS estimator is characterised by poor efficiency compared to other estimators (Rousseeuw, 1984), i.e. it is characterised by a slow convergence rate. Furthermore, since it is not based on the minimisation of the sum of squared residuals, standard inference procedures cannot be directly applied.

To overcome these limitations, a robust weighted least squares estimator was developed by Rousseeuw and Leroy (1987) where binary weights are introduced in the objective function to include or exclude measurements from the fitting. The estimation of the parameters requires the solution of a Mixed-Integer NonLinear Program (MINLP) and retains the highest breakdown point theoretically achievable. The approach was also extended to multi-response models (Hubert et al., 2008) with the idea of detecting outlying samples involving multiple measurements. An interesting feature of the robust weighted least squares estimator is that it can be employed to perform unsupervised model-based data mining (MBDM) and effectively classify samples in terms of good or bad model predictive performance. Let $\beta_i \in \{+1,-1\}\ \forall\, i = 1,...,N$ be binary weights. The robust weighted least squares approach proposed by Rousseeuw et al. (2004) can be formulated as follows where the objective function to maximise is a modified log-likelihood function $\mathscr{L}_{DM}$ given in (2.28).

$$\hat{\theta}_{DM} = \arg\max_{\theta\in\Theta} \mathscr{L}_{DM} \tag{2.27}$$

$$\mathscr{L}_{DM} = \sum_{i=1}^{N} \frac{1+\beta_i}{2} \cdot \{N_y c^2 - [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]\} \tag{2.28}$$

$$\text{s.t.} \quad \beta_i(\theta) = \begin{cases} +1 & \text{if} \quad [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)] \leq N_y c^2 \\ -1 & \text{if} \quad [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)] > N_y c^2 \end{cases} \quad \forall \quad i \tag{2.29}$$

The binary weights are introduced to control the inclusion ($\beta_i = +1$) or exclusion ($\beta_i = -1$) of samples in the objective function $\mathscr{L}_{DM}$. The binary weights are subject to the conditions in (2.29). In words, the conditions in (2.29) state that if the residuals associated with the i-th sample (i.e. $\mathbf{y}_i$) are too large, then the sample $\mathbf{y}_i$ is excluded from the objective function and ignored for the fitting of the model parameters. The quantity c appearing in (2.28) and (2.29) is a user-defined hyperparameter which quantifies the maximum threshold

of acceptance for a model residual. The value of c shall be set ≥ 2.0 to reduce the chance of excluding samples whose residuals are compatible with measurement noise (Rousseeuw and Leroy, 1987).

The quantity $\hat{\theta}_{DM}$ represents a robust maximum likelihood estimate obtained from the fitting of a possibly reduced dataset

$$Y' = \{\mathbf{y}_i \in Y \,|\, \beta_i(\hat{\theta}_{DM}) = +1\} \subseteq Y \tag{2.30}$$

The covariance associated with the estimate $\hat{\theta}_{DM}$ can be obtained from the observed information matrix constructed using the reduced dataset Y' as follows

$$\mathbf{V}_\theta = [-\nabla\nabla^T \mathscr{L}(Y'|\hat{\theta}_{DM})]^{-1} \tag{2.31}$$

where $\mathscr{L}(Y'|\hat{\theta}_{DM})$ is the log-likelihood function constructed on the reduced dataset Y' evaluated at $\theta = \hat{\theta}_{DM}$.

2.7 Model structure improvement

In likelihood-based inference, model parameters are estimated by maximising the likelihood function (or equivalently the log-likelihood function) and the modelling hypotheses are checked using a goodness-of-fit test (Silvey, 1975). The two-tailed goodness-of-fit test illustrated in Section 2.4.4.1 can inform on the appropriateness of the model in representing the data. Nonetheless, whenever over-fitting or under-fitting is detected by a failed goodness-of-fit test, no information is obtained on how the model structure can be improved.

In general, in parametric modelling, it is desirable that the number of free model parameters N_θ is kept as small as possible. The inclusion of unnecessary parameters in the model (e.g. the inclusion of an additional reaction that is not actually taking place in the system) typically causes an increase in the confidence range of the parameter estimates given the same dataset. As a consequence, more experimental resources will be required to obtain estimates with the desired level of statistical quality. Statistical tests were proposed in the literature to assess whether it is possible to refine a model by reducing its number of free parameters, i.e. by applying constraints on the parameters, without causing a significant degradation of the model fitting quality. More specifically, the likelihood ratio (LR) test, the Wald (W) test and the Lagrange multipliers (LM) test (Buse, 1982) are regularly applied for structural equation modelling in many applied sciences including psychometrics and econo-

metrics (Green et al., 1999; Engle, 1984; Chou and Bentler, 1990; Anselin, 1988). These tests were proposed to evaluate the generic null hypothesis that parameters satisfy a certain set of $N_s < N_\theta$ constraints

$$\mathbf{s}(\theta) = \mathbf{0} \tag{2.32}$$

where $\mathbf{s}$ is a $N_s \times 1$ array of functions of the model parameters. As an example, one may use the aforementioned tests to challenge the hypothesis that some parameters are equal to zero. If there is not sufficient evidence from the data for disproving this hypothesis, the considered parameters should be fixed to zero and treated as constants.

The tests are asymptotically equivalent and assume the same null hypothesis, but they differ significantly in the construction of their test statistics (Chandra and Joshi, 1983). In the following, the symbol $\hat{\theta}$ is used to denote the ML estimate obtained by maximising the unconstrained log-likelihood function as in (2.8). Let $\hat{\theta}_s$ be the ML estimate obtained by maximising the log-likelihood function under the constraints (2.32). The ML estimate $\hat{\theta}_s$ satisfies the constrained ML equations

$$\begin{aligned} \nabla \mathscr{L}(Y|\hat{\theta}_s) + \nabla \mathbf{s}(\hat{\theta}_s)\hat{\alpha} &= \mathbf{0} \\ \mathbf{s}(\hat{\theta}_s) &= \mathbf{0} \end{aligned} \tag{2.33}$$

where $\hat{\alpha}$ denotes a $N_s \times 1$ array of Lagrange multipliers.

The Wald statistic ξ_W is defined as

$$\xi_W = \mathbf{s}(\hat{\theta})^T \left[\nabla \mathbf{s}(\hat{\theta})^T \mathbf{V}_\theta \nabla \mathbf{s}(\hat{\theta})\right]^{-1} \mathbf{s}(\hat{\theta}) \tag{2.34}$$

and it is a measure of the distance between the unconstrained and the constrained maximum likelihood estimates in the parameter space (Wald, 1943).

The Lagrange multipliers statistic is computed as

$$\xi_{LM} = \hat{\alpha}^T \nabla \mathbf{s}(\hat{\theta}_s)^T \mathbf{H}(\hat{\theta}_s)^{-1} \nabla \mathbf{s}(\hat{\theta}_s)\hat{\alpha} \tag{2.35}$$

and it is a function of the log-likelihood gradient at the constrained estimate (Silvey, 1959; Bera and Bilias, 2001; Rao, 1948).

The statistic used in the likelihood ratio test is

$$\xi_{LR} = 2\left[\mathscr{L}(Y|\hat{\theta}) - \mathscr{L}(Y|\hat{\theta}_s)\right] = \chi^2_Y(\hat{\theta}) - \chi^2_Y(\hat{\theta}_s) \tag{2.36}$$

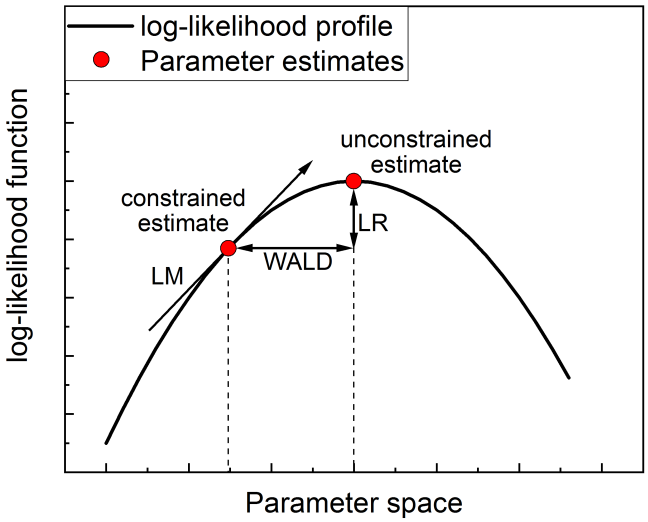


Figure 2.7: A graphical interpretation of the Likelihood Ratio (LR), Wald and Lagrange Multiplier (LM) statistics for hypothesis testing in model identification frameworks based on maximum likelihood inference. If the statistics are *small*, the presence of the constraint is not disproved and the constrained model should be selected because it involves a smaller number of parameters.

and quantifies the distance between the constrained and the unconstrained estimates in terms of log-likelihood values (Wilks, 1938). A graphical interpretation of the statistics for the different tests is given in Figure 2.7.

All the above statistics are asymptotically distributed as a χ^2 distribution with degree of freedom N_s under the null hypothesis being true (Engle, 1984). However, depending on the specific case, one statistic may be significantly more convenient to compute than the others (Engle, 1984). In fact, while the likelihood ratio test requires both the constrained and the unconstrained estimates to be computed, the Wald test and Lagrange multipliers test require respectively only the unconstrained and the constrained estimates.

Whenever a proposed kinetic model is falsified for over-fitting, one may regard its parameter estimates as an unconstrained estimate. A Wald test may then inform on which parameters can be constrained to zero and removed from the model structure. When under-fitting is detected, a change in the model structure may be required. An approach for improving an under-fitting model consists of regarding the proposed model as a constrained instance of one or multiple alternative superstructures (Breusch and Pagan, 1980; Engle, 1982). A Lagrange multipliers test may then be performed to challenge the constrained

model against the available alternatives without the necessity or re-estimating the model parameters for each superstructure. A limitation of this approach is that the definition of appropriate superstructures still relies entirely on the intuition of the modeller.

2.8 Summary of literature review

A survey of the scientific literature on the field of parametric modelling for describing chemical kinetics was presented in this Chapter. In Section 2.2, kinetic models were classified as *knowledge-driven*, *data-driven* and *hybrid*, highlighting the fact that systematic techniques for the identification of hybrid models were not pursued as systematically as for the previous two categories. As a consequence, whenever the knowledge available on the system behaviour is not sufficient to build a knowledge-driven model, an entirely data-driven approach is typically adopted and any valuable information on the system mechanisms is completely neglected.

Techniques from the literature on statistical inference were presented in Section 2.4 for bridging modelling and experimentation in kinetic modelling studies. It was shown that the problem of inferring the values of kinetic model parameters from input-output experiments requires that the model structure satisfies requirements for *a-priori* identifiability (see Section 2.4.2). Nonetheless, it was also highlighted that *a-priori* identifiability is only a necessary, but not sufficient, condition for model-based inference and that it may still be impossible or extremely challenging to estimate model parameters from noisy experimental data. In Section 2.5, a survey of the techniques available for addressing practical identifiability problems is presented. Techniques were classified as *experimental-design-based* (ED), *regularisation-based* (RG) and *reparametrisation-based* (RP). It was shown that, in contrast to ED-based and RG-based methods, RP-based methods do not require the execution of experiments nor the introduction of bias in the parameter estimation problem to address practical identifiability issues. Nonetheless, it was also observed that general RP-based approaches have received little attention from the scientific community and RP-based methods have never been considered in the context of online kinetic modelling studies.

In Section 2.4, it was also shown that standard techniques for parameter estimation and model-based experimental design account for the measurement noise present in the system, but do not account for the possible systematic deviations between observations and model predictions. Data deviating from the assumptions are called outliers and their presence in the dataset may be associated with a number of causes: from experimental disturbances to

systematic errors and inappropriate modelling assumptions. A review of robust regression methods for parameter estimation in the presence of outlier contamination in the dataset was given in Section 2.6. It is observed that robust regression methodologies are seldom employed in kinetic modelling practice. Furthermore, the development of criteria for optimal experimental design that are robust towards systematic errors and/or model misspecification were not pursued as systematically as robust regression methods.

Statistical tools for refining model structures in likelihood-based modelling frameworks were discussed in Section 2.7. It is shown that a number of statistical tests, i.e. the *likelihood-ratio* test, the *Wald* test and the *Lagrange multipliers* test, are available in the literature to support the scientist in the identification of appropriate parameter constraints to reduce the number of free parameters in a model without causing a significant degradation of the model fitting quality. These tests can be directly applied to remove irrelevant parameters whenever a candidate model is over-fitting. Nonetheless, few systematic approaches have been proposed in the literature to support the scientist in the improvement of under-fitting models, which is a task that still relies almost entirely on the intuition of experienced researchers.

The work presented in the following research Chapters aims at bridging the aforementioned gaps in the literature. An online-RP approach is proposed in Chapter 3 to systematically reparametrise the model equations in the course of online kinetic modelling studies. A framework for the estimation of parameters in approximated kinetic models is introduced in Section 4, where a criterion for optimal MBDoE in the presence of model misspecification is also proposed. A statistical test based on maximum likelihood inference is formulated in Chapter 5 for diagnosing model misspecification in under-fitting models, i.e. in the presence of significant process-model mismatch. In Chapter 6, further tests are formulated to support the scientist in the improvement of under-fitting model structures. The content of the following Chapters is also summarised in the Thesis roadmap in Figure 1.3.

Chapter 3

Online model reparametrisation for robust parameter estimation

Part of this Chapter is adapted from the following articles:

Quaglio M., Waldron C., Pankajakshan A., Cao E., Gavriilidis, A., Fraga E. S., Galvanin F., On the use of online reparametrization in automated platforms for kinetic model identification, *Chemie Ingenieur Technik* 91(3), 2019, pp. 268-276

Quaglio M., Waldron C., Pankajakshan A., Cao E., Gavriilidis A., Fraga E. S., Galvanin F., An online reparametrisation approach for robust parameter estimation in automated model identification platforms, *Computers & Chemical Engineering* 124, 2019, pp. 270-284

The author of this Thesis contributed to the above articles by developing the main novel ideas, implementing the simulations, and writing a significant part of the text. Hence, the author retains the right to include the articles in this Thesis since it is not published commercially and the journals are referenced as the original source.

3.1 Introduction

The structure of kinetic models is frequently affected by problems of practical identifiability. In different words, the fitting quality may be insensitive to a change in the parameter values and/or parameters may be extremely correlated. Whenever the model exhibits this type of behaviour it is called sloppy (see Section 2.5) (White et al., 2016). Standard model identification algorithms are prone to numerical failures in the presence of a sloppy parametrisation (see Section 1.2.2.1). Numerical failures may result in the invalidation of the model identification process and a significant waste of experimental resources, espe-

cially if the identification of the model is performed online on an autonomous setup without scientist supervision. An approach for the identification of sloppy models in online model identification platforms is presented in this Chapter. The central step in the framework is an automated approach to online model reparametrisation. The goal of online reparametrisation is to control model sloppiness by optimally transforming the parameter space every time new samples become available to the online model identification algorithm. Parameter estimation and MBDoE problems are solved in a transformed, robust parameter space where the risk of numerical failures is low. The approach is demonstrated both in-silico and in a closed-loop system on the identification of a kinetic model of catalytic esterification of benzoic acid with ethanol in an automated flow microreactor system.

3.2 Proposed methodology

It is assumed that an experimental platform for kinetic model identification is available to study the dynamics of a chemical process of interest. A set of N_y physical quantities can be sampled in the experiments. The sample is denoted with the $N_y \times 1$ vector $\mathbf{y}$. Measurements of $\mathbf{y}$ are affected by Gaussian noise with zero mean and covariance Σ_y. A preliminary dataset Y is available and it consists of N samples of $\mathbf{y}$, i.e., $Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$. The scientist proposes a model structure in the general form (2.1) to describe the dynamics of the system

$$\begin{aligned} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t, \theta) &= \mathbf{0} \\ \hat{\mathbf{y}} &= \mathbf{h}(\mathbf{x}, \mathbf{u}, t, \theta) \end{aligned} \tag{2.1}$$

An online approach to model reparametrisation is now introduced with the aim of effectively estimating the parameter set $\theta \in \Theta$. The original set of model equations is initially extended including a linear system of equations to transform the parameter space.

$$\theta = \mathbf{G}\omega \tag{3.1}$$

In (3.1), $\omega \in \Omega$ represents the $N_\theta \times 1$ array of model parameters in the transformed parameter space Ω, $\mathbf{G}$ is a $N_\theta \times N_\theta$ matrix which transforms the parameter space Ω to the original model parameter space Θ. A diagram showing the proposed procedure is given in Figure 3.1. The parameter transformation $\mathbf{G}$ is initially set equal to $\mathbf{I}_\theta$, where $\mathbf{I}_\theta$ is the $N_\theta \times N_\theta$ identity matrix, i.e., the parameter spaces Θ and Ω are initially coincident. The model identification algorithm is then called providing the available dataset as input. The fundamental

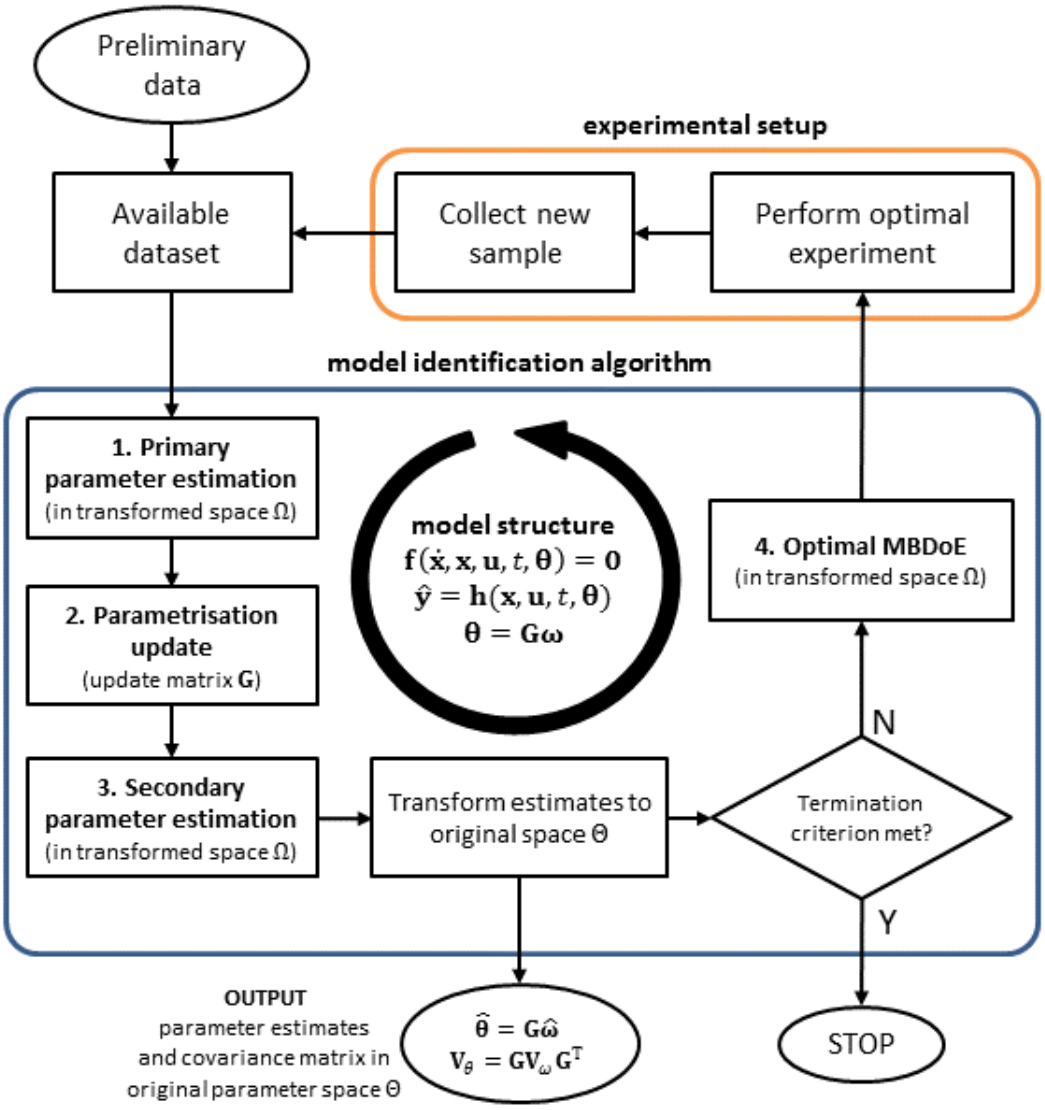


Figure 3.1: Proposed framework for the online identification of sloppy models. Fundamental step in the procedure is the update of the parametrisation matrix **G** after the collection and fitting of each sample. The online modification of the model parametrisation is performed to maintain a high computational performance at the parameter estimation and optimal MBDoE stages in the procedure.

steps in the algorithm are:

1. *A primary parameter estimation step.* At this stage, the set of transformed parameters $\boldsymbol{\omega}$ is estimated fitting the available dataset using a maximum likelihood approach (Bard, 1974). The Hessian of the likelihood function is then computed to characterise the geometry of the parameter space and quantify its *sloppiness*.

2. *A parametrisation update step.* The Hessian matrix computed at the *primary parameter estimation* step is employed to compute an update for the transformation matrix

$\mathbf{G}$ with the aim of minimising the condition number (i.e., eliminating the sloppiness) given the available dataset.

3. *A secondary parameter estimation step.* The model parameters $\boldsymbol{\omega} \in \Omega$ are estimated after the *parametrisation update* step and their statistical quality is quantified by computing their covariance matrix $\mathbf{V}_{\omega}$. Parameter estimates and related covariance computed in the transformed parameter space Ω are then transformed to the original parameter space Θ and returned as output.

4. *An optimal MBDoE for parameter precision step.* If parameter statistics in Θ are unsatisfactory and the experimental budget allows for additional samples to be collected, the experimental activity shall proceed. Optimal experimental conditions for the collection of additional samples are identified at this stage through MBDoE techniques for parameter precision (Franceschini and Macchietto, 2008b). The optimal MBDoE problem is solved in the transformed parameter space Ω.

The illustrated steps constitute an iteration in the presented online framework. These are further detailed in the following subsections. The computational burden associated with the application of the proposed methodology is comparable with standard parameter estimation algorithms based on parameter fitting. The procedure shows how it is possible to achieve an effective estimation of parameters in a (potentially) sloppy parameter space Θ by invoking the parameter estimation and the MBDoE algorithms in a conveniently transformed, non-sloppy, parameter space Ω. Operations of optimisation and matrix inversion are performed in the robust parameter space Ω where the risk of numerical failures is low. The values of the estimates and the related covariance obtained in Ω are then transformed to the original parameter space Θ by applying linear transformations.

3.2.1 Primary parameter estimation

The available dataset Y is provided to the model identification algorithm (see Figure 3.1). The transformation matrix $\mathbf{G}$ is set equal to the *primary* transformation matrix $\mathbf{G}_{\mathrm{P}}$. At the beginning of the model identification procedure $\mathbf{G}_{\mathrm{P}}$ is initialised as the identity matrix $\mathbf{I}_{\theta}$. A primary estimation of the model parameters $\hat{\boldsymbol{\omega}}_{\mathrm{P}}$ is performed as in (3.3) maximising the

log-likelihood function (3.2).

$$\begin{aligned}\mathscr{L}(Y|\boldsymbol{\omega})|_{\mathbf{G}=\mathbf{G}_\mathrm{P}} = &-\frac{N}{2}[N_y \ln(2\pi) + \ln(\det(\Sigma_y))] \\ &-\frac{1}{2}\sum_{i=1}^{N}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\boldsymbol{\omega})]^T \Sigma_y^{-1}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\boldsymbol{\omega})]|_{\mathbf{G}=\mathbf{G}_\mathrm{P}}\end{aligned} \tag{3.2}$$

$$\hat{\boldsymbol{\omega}}_\mathrm{P} = \arg\max_{\boldsymbol{\omega}\in\Omega} \mathscr{L}(Y|\boldsymbol{\omega})|_{\mathbf{G}=\mathbf{G}_\mathrm{P}} \tag{3.3}$$

In (3.2), the quantity $\hat{\mathbf{y}}_i$ represents the model prediction for the sample $\mathbf{y}_i$. The negative Hessian $\mathbf{H}$ of the log-likelihood function is then computed to evaluate the geometrical properties of the log-likelihood profile in proximity of the maximum likelihood estimate as

$$\mathbf{H}(\hat{\boldsymbol{\omega}}_\mathrm{P})|_{\mathbf{G}=\mathbf{G}_\mathrm{P}} = -\nabla\nabla^T \mathscr{L}(Y|\hat{\boldsymbol{\omega}}_\mathrm{P})|_{\mathbf{G}=\mathbf{G}_\mathrm{P}} \tag{3.4}$$

In (3.4), the symbol ∇ defines the gradient operator in the parameter space Ω. Matrix $\mathbf{H}$ is also known as the observed Fisher information matrix and its inverse quantifies the covariance matrix of the parameter estimates (Pukelsheim, 2006).

3.2.2 Parametrisation update

An eigendecomposition of the matrix (3.4) is performed at this stage with the aim of diagnosing the structure of the log-likelihood function in proximity of the maximum likelihood estimate and compute an opportune update to the transformation matrix $\mathbf{G}$. Let Λ be the diagonal matrix whose diagonal elements are the eigenvalues $\lambda_1, ..., \lambda_{N_\theta}$ of the observed Fisher information matrix (3.4). The eigenvalues of the observed Fisher information matrix represent the inverse eigenvalues of the parameter covariance matrix and the ratio between the maximum and the minimum eigenvalue is the condition number κ.

$$\kappa = \frac{\max_i \lambda_i}{\min_i \lambda_i} \tag{2.24}$$

Let matrix $\mathbf{U}$ be the matrix whose columns represent the right normalised eigenvectors of the observed Fisher information matrix (3.4). Matrix Λ and matrix $\mathbf{U}$ quantify the sloppiness of the model in a more readable format. In fact, the eigenvalues and eigenvectors of the negative Hessian (3.4) respectively quantify the extent of the sloppiness and the directions in the parameter space which are associated to the sloppy behaviour of the model (López Cárdenas et al., 2015). A family of *secondary* transformations $\mathbf{G}_\mathrm{S}$ can be

constructed from $\mathbf{G}_\mathrm{P}$, $\mathbf{U}$ and Λ as in (3.5) with the aim of minimising the condition number of the problem (i.e., making $\kappa = 1.0$).

$$\mathbf{G}_\mathrm{S} = d\,\mathbf{G}_\mathrm{P}\mathbf{U}\Lambda^{-\frac{1}{2}}\mathbf{R} \tag{3.5}$$

The family of transformations given in (3.5) is parametrised by the scalar $d > 0$ and by the matrix $\mathbf{R}$, which represent respectively a scaling factor and a rotation matrix in the parameter space. The condition number κ is not influenced by the choice of d and $\mathbf{R}$. However, the omission of d and $\mathbf{R}$ from (3.5) (the omission is equivalent to setting $d = 1.0$ and $\mathbf{R} = \mathbf{I}_\theta$) may result in a transformation to a new parameter space in which there is significant discrepancy in the orders of magnitude of the model parameters. Model identification algorithms are influenced by the relative scale of parameters, e.g. in the computation of the gradients and, consequently, in the computation of the covariance of parameter estimates (Saltelli et al., 2000). Working with parameters sharing the same order of magnitude is therefore desirable to avoid discrepancies on how the model identification algorithm handles different directions of the parameter space. In this work, the scaling factor d and the matrix $\mathbf{R}$ are computed to scale the parameter values to the same order of magnitude.

The secondary transformation matrix $\mathbf{G}_\mathrm{S}$, computed as in (3.5), is then used to *update* the primary transformation matrix $\mathbf{G}_\mathrm{P}$ that will be used at the following iteration in the procedure of Figure 3.1.

3.2.3 Secondary parameter estimation

The aim at the *secondary parameter estimation* stage is to obtain a more accurate estimate for the parameters in the transformed space Ω. This is done by repeating the estimation of the parameters $\boldsymbol{\omega}$ after the *parametrisation update* stage, i.e., after the transformation of the (possibly) sloppy parameter space in a more robust, non sloppy parameter space. The log-likelihood function of the model is optimised as in (3.6) with $\mathbf{G} = \mathbf{G}_\mathrm{S}$ obtaining the *secondary* parameter estimate $\hat{\boldsymbol{\omega}}_\mathrm{S}$.

$$\hat{\boldsymbol{\omega}}_\mathrm{S} = \arg\max_{\boldsymbol{\omega}\in\Omega} \mathscr{L}(Y|\boldsymbol{\omega})|_{\mathbf{G}=\mathbf{G}_\mathrm{S}} \tag{3.6}$$

In principle, the *primary* and the *secondary* parameter estimates satisfy the equality $\mathbf{G}_\mathrm{P}\hat{\boldsymbol{\omega}}_\mathrm{P} = \mathbf{G}_\mathrm{S}\hat{\boldsymbol{\omega}}_\mathrm{S}$. However, numerical algorithms for parameter estimation are sensitive to the model parametrisation (Rimensberger and Rippin, 1986; Dovi et al., 1994). More

specifically, the convergence rate of numerical optimisation routines to the maximum likelihood estimate is sensitive to the choice of the transformation matrix $\mathbf{G}$ and the aforementioned equality may not be satisfied in practice (Higham, 1996). The covariance $\mathbf{V}_\omega$ is then computed for the *secondary* parameter estimates as the inverse of the observed Fisher information matrix (Bard, 1974).

$$\mathbf{V}_\omega = \left[\mathbf{H}(\hat{\boldsymbol{\omega}}_S)|_{\mathbf{G}=\mathbf{G}_S}\right]^{-1} \tag{3.7}$$

The parameter estimates $\hat{\theta}$ and their associated covariance matrix $\mathbf{V}_\theta$ in the original parameter space Θ are then computed by applying the *secondary* transformation to the estimates $\hat{\boldsymbol{\omega}}_S$ and covariance $\mathbf{V}_\omega$ computed in the transformed space Ω.

$$\hat{\theta} = \mathbf{G}_S \hat{\boldsymbol{\omega}}_S \tag{3.8}$$

$$\mathbf{V}_\theta = \mathbf{G}_S \mathbf{V}_\omega \mathbf{G}_S^T \tag{3.9}$$

In standard parameter estimation algorithms, the computation of the covariance $\mathbf{V}_\theta$ requires the inversion of the information matrix in the original parameter space Θ (Bard, 1974). However, in the presence of a sloppy parametrisation, the information matrix in Θ may be ill-conditioned. Notice that, in the proposed framework, the inversion of ill-conditioned matrices is avoided. In fact, matrix inversion is performed in a conveniently transformed parameter space Ω, as in (3.7), where the information matrix is well-conditioned. The covariance in the original parameter space $\mathbf{V}_\theta$ is then computed as in (3.9) by applying algebraic transformations, which are numerically more robust operations than matrix inversions (Higham, 1996).

3.2.4 Optimal MBDoE for parameter precision

If some parameter statistics are not satisfactory and the experimental budget allows for the collection of additional data then the experimental activity will continue with the collection of an additional sample from the experimental setup. The following sample will be collected with the aim of minimising the size of the confidence region of the parameter estimates $\hat{\theta} \in \Theta$.

Optimal MBDoE problems for parameter precision may be ill-conditioned in the presence of a sloppy parametrisation (White et al., 2016). In fact, the solution of an optimal

MBDoE problem requires the inversion of an ill-conditioned matrix if the parametrisation is sloppy. In this work it is proposed to solve the MBDoE problem in the robust parameter space Ω with the aim of minimising the size of the confidence region in the original parameter space Θ. In general, the optimal experimental conditions depend on the type of criterion adopted for the design (see Section 2.4.6 for more information on experimental design criteria) and on the model parametrisation. In this study, the D-optimal criterion is employed because it is invariant under linear transformations of the parameter space (Fedorov, 1972; Rimensberger and Rippin, 1986). In fact, the following equality holds:

$$\det(\mathbf{V}_\theta) = \det(\mathbf{G}_\mathrm{S})^2 \det(\mathbf{V}_\omega) \tag{3.10}$$

It is sufficient to notice that matrix $\mathbf{G}_\mathrm{S}$ is not modified at the *optimal MBDoE* stage of the procedure (see Figure 3.1), i.e., $\det(\mathbf{G}_\mathrm{S})$ represents a constant in the MBDoE problem. Therefore, minimising the determinant of the covariance $\det(\mathbf{V}_\omega)$ in the transformed parameter space Ω is equivalent to minimising the determinant of the covariance $\det(\mathbf{V}_\theta)$ in the original parameter space Θ.

The optimal MBDoE problem in the robust space Ω requires the computation of a prediction for the parameter covariance $\hat{\mathbf{V}}_\omega$ (i.e., the posterior covariance matrix) after the collection of the new sample.

$$\hat{\mathbf{V}}_\omega = \left[\mathbf{V}_\omega^{-1} + \nabla\hat{\mathbf{y}}(\hat{\boldsymbol{\omega}}_\mathrm{S})\Sigma_y^{-1}\nabla\hat{\mathbf{y}}(\hat{\boldsymbol{\omega}}_\mathrm{S})^T|_{\mathbf{G}=\mathbf{G}_\mathrm{S}}\right]^{-1} \tag{3.11}$$

In (3.11), the second addend in the bracket represents the expected Fisher information matrix of the sample to be designed, which is a function of the experimental design vector $\boldsymbol{\varphi}$. The inverse of the prior covariance matrix $\mathbf{V}_\omega$ is also included in (3.11) to quantify the preliminary information that is available from previously fitted samples. The prior covariance is updated at every iteration of the procedure in Figure 3.1, i.e., after the collection of each sample, according to (3.7). The D-optimal experimental conditions $\boldsymbol{\varphi}^*$ for the collection of the following sample are computed solving the following optimisation problem

$$\boldsymbol{\varphi}^* = \arg\min_{\boldsymbol{\varphi}\in\Phi} \det(\hat{\mathbf{V}}_\omega) \tag{3.12}$$

3.3 Case study

The proposed algorithm presented in Section 3.2 is integrated in an automated platform for kinetic model identification and tested on a case study. The objective is the identification of a kinetic model of benzoic acid esterification with ethanol in a microreactor system (Pipus et al., 2000). The reaction is homogeneous and it is catalysed by sulphuric acid. A description of the automated model identification platform is given in Section 3.3.1. The modelling assumptions are presented in Section 3.3.2. The proposed online RP methodology is tested both in-silico (Section 3.4.1) and experimentally on the automated system (Section 3.4.2). For both the simulated and the real cases two experimental campaigns are performed:

- a campaign where the parametrisation matrix is not modified (non-RP campaign);
- a campaign where the parametrisation matrix is updated online (RP campaign).

The two campaigns are performed to assess the influence of the online RP on the model identification process. The methods adopted for the conduction of the experimental campaigns are detailed in Section 3.3.3.

3.3.1 Automated model identification platform

A simplified diagram for the online model identification platform is given in Figure 3.2. The esterification of benzoic acid with ethanol catalysed by sulphuric acid occurs in a flow microreactor. The microreactor is a 2 m long PEEK tube with a diameter of 250 μm. It is placed in a stirred oil bath whose temperature is controlled by a rope heater. The reactants and the catalyst are injected through the flow reactor by three syringe pumps. Syringe 1 and syringe 2 are filled with two different mixtures of benzoic acid and ethanol. The feed concentration of benzoic acid in the reactor is manipulated by modifying the relative flowrates of syringe 1 and syringe 2. Syringe pump 3 is filled with a 160 g L^{-1} sulphuric acid solution. The flowrate of syringe 3 is kept at 10% of the overall flowrate to maintain a constant concentration of sulphuric acid at 16 g L^{-1} at the inlet of the flow reactor. The mixture at the outlet of the reactor is analysed online by a Jasco HPLC using a 250 mm long, 4.6 mm internal diameter ODS hypersil column with a particle size of 5 μm (Thermo Fisher Scientific).

The experimental conditions which can be manipulated by the automated system are:

- the inlet concentration of benzoic acid C_{BA}^{IN} in the range 0.9 - 1.55 mol L^{-1};

- the flowrate F of the feed mixture to the reactor in the range 7.5 - 30.0 μL min^{-1};
- the temperature of the oil bath T in the range 343.0 - 413.0 K.

These constitute independent directions of the explorable space of experimental conditions $\Phi = (C_{BA}^{IN}, F, T)$. The experimental setup is controlled through a LabVIEW interface (Elliott et al., 2007) implemented in a 32-bit Windows machine with Intel® Core® i7-3770 3.40 GHz processor and 4.0 GB of RAM. A script written in Python 2.7 implementing the model identification algorithm presented in Section 3.2 is integrated with LabVIEW for the purposes of online parameter estimation and sample design. The main Python packages employed in the script are NumPy 1.13 (Oliphant, 2015) for the manipulation of algebraic objects and SciPy 1.1 (Jones et al., 2001) for integrating the kinetic model equations and solving the optimisation problems associated with parameter estimation and MBDoE. Parameter estimation problems are solved using the *Nelder-Mead* method. MBDoE problems are solved employing the *SLSQP* solver.

The *parametrisation update* stage of the algorithm (see Figure 3.1) was implemented in the Python script as an option that can be activated or deactivated from LabVIEW. This option was implemented to give more flexibility to the user in testing the model identification algorithm both in the presence and in the absence of the online RP method.

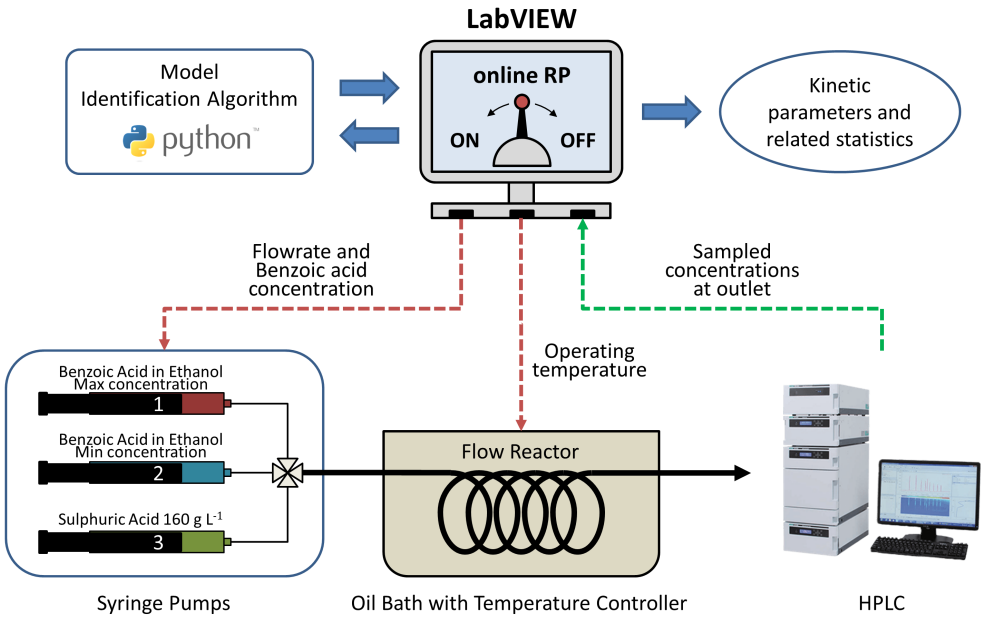


Figure 3.2: Simplified diagram representing the online model identification platform.

3.3.2 Modelling assumptions

The catalytic esterification of benzoic acid and ethanol is modelled as a single reaction system where benzoic acid (BA) and ethanol (Et) react to produce ethyl benzoate (EB) and water (W) (Pipus et al., 2000).

$$\text{Benzoic Acid} + \text{Ethanol} \leftrightharpoons \text{Ethyl Benzoate} + \text{Water} \tag{3.13}$$

Available studies in the literature report that the reaction is reversible. However, if a large excess of ethanol in the reactor is maintained (as in this work), the reverse reaction can be neglected (Pipus et al., 2000). The tubular reactor is modelled as an ideal plug flow reactor operated at isothermal conditions, i.e., thermal and mass transfer resistances are neglected. The validity of plug flow behaviour was checked by evaluating the vessel dispersion number (Levenspiel, 1998; Rossi et al., 2017). A maximum vessel dispersion number of $6.8 \cdot 10^{-4}$ was computed for the flowrate range considered in the study. The computed value is significantly smaller than $1.28 \cdot 10^{-2}$, i.e., the maximum vessel dispersion number recommended in the literature for the validity of the plug flow assumption (Levenspiel, 1998).

The reaction rate is assumed as first order with respect to benzoic acid. Following from the aforementioned assumptions, the steady-state kinetic behaviour of the system is modelled through the following set of ordinary differential equations (3.14):

$$v\frac{dC_i}{dz} = \nu_i k C_{\text{BA}}(z) \quad \forall\, i = \text{BA}, \text{Et}, \text{EB}, \text{W} \tag{3.14}$$

In (3.14), C_i is the concentration of the i-th component in the mixture expressed in $\text{mol}\,\text{L}^{-1}$; z represents the axial spatial coordinate of the tubular reactor expressed in m; v is the axial velocity of the liquid bulk expressed in $\text{m}\,\text{s}^{-1}$; ν_i is the stoichiometric coefficient of the i-th component in the mixture; k is the rate constant expressed in s^{-1}.

An Arrhenius-type kinetic constant involving a set of two kinetic parameters $\theta = [\theta_1, \theta_2]$ is assumed with the following mathematical structure:

$$k = e^{\theta_1 - \frac{10^4 \theta_2}{RT}} \tag{3.15}$$

In (3.15), R is the ideal gas constant expressed in $\text{J}\,\text{mol}^{-1}\,\text{K}^{-1}$. As one can see from (3.15), the pre-exponential factor is included as exponent in the rate constant and the activa-

tion energy is multiplied by a scaling factor. The above structure for the kinetic rate constant was selected because it is generally recognised as robust within the literature on kinetic parameter estimation (Asprey and Naka, 1999; Buzzi-Ferraris and Manenti, 2009). In other words, parametrisation (3.15) generally leads to an improvement of the condition number with respect to the original form of the Arrhenius constant, i.e. $k = Ae^{-E_a/RT}$, parametrised by pre-exponential factor A and activation energy E_a.

3.3.3 Objective and methods

The objective of the study is the estimation of the kinetic parameters $\theta = [\theta_1, \theta_2]$ with the smallest volume confidence region of $\hat{\theta}$ by conducting an experimental campaign with an available budget of 9 samples. A sample is constituted by the single measurement of ethyl benzoate concentration at the outlet of the reactor, i.e., $\mathbf{y} = [C_{\text{EB}}^{\text{OUT}}]$ mol L^{-1}. The measurement error is modelled as Gaussian noise with covariance matrix $\Sigma_y = [2.5 \cdot 10^{-5}]$, i.e., a standard deviation of 0.0165 mol L^{-1} is assumed to model the Gaussian measurement noise for $C_{\text{EB}}^{\text{OUT}}$. The experimental conditions for the collection of samples 1, 2 and 3 are fixed to the values reported in Table 3.1. The following samples, i.e., samples from 4 to 9, are designed by the model identification algorithm by employing a D-optimal criterion, i.e., by solving an MBDoE problem in the form (3.12).

Two cases are proposed to test the model identification algorithm implemented in the online model identification platform:

1. *Simulated case: samples generated in-silico.* Samples are generated simulating the experiments with the kinetic model (3.14) setting the kinetic parameters equal to the value $\theta^* = [15.27, 7.60]$ and adding Gaussian noise with covariance Σ_y.

2. *Real case: samples collected from the experimental platform.* In this case, samples are collected from the experimental platform described in Section 3.3.1. An interval of 65 min is allowed between the collection of samples to let the system reach steady-state conditions.

For both the *Simulated* and the *Real* case, two experimental campaigns are performed: 1) a non-RP campaign in which the online reparametrisation is not activated; 2) a RP campaign in which the online reparametrisation is activated. This is done to provide a comparison of the performance of the model identification algorithm both in the presence and in the absence of the online RP method. In the *Simulated* case, the effect of the online RP is

assessed comparing statistically the parameter estimates $\hat{\theta}$ computed in the two campaigns with the target parameter value $\theta^* = [15.27, 7.60]^T$. This is done by means of a χ^2-test in the parameter space Θ. This involves testing the null hypothesis that the following statistic χ^2_θ is distributed as a χ^2 distribution with degree of freedom $N_\theta = 2$.

$$[\hat{\theta} - \theta^*]^T \mathbf{V}_\theta^{-1} [\hat{\theta} - \theta^*] = \chi^2_\theta \sim \chi^2 \tag{3.16}$$

A small p-value associated to the statistic χ^2_θ (e.g. smaller than 1.0%) is interpreted as an index of failure of the model identification algorithm in estimating the target parameter values. In the *Real* case, the target parameter value θ^* is unknown. Furthermore, a discrepancy in the parameter estimates between the RP and the non-RP campaigns is not only caused by numerical reasons, but also by problems of experimental repeatability caused by external disturbances (Alberton et al., 2009). The presence of disturbances can lead to changes in the parameters of the population from which experimental data are sampled and the concomitant inclusion of outliers in the dataset (Huber, 2004). It is recognised that, in the presence of such uncertainty sources, a statistical analysis to validate the models identified in the two campaigns would not be significant and it is therefore omitted.

The confidence intervals and the correlation coefficient c_{12} associated with the estimates (see Section 2.4.4.2) are recorded in the course of the experimental campaigns and they are reported in Section 3.4. The condition number κ is also recorded in the course of the experimental campaigns and it is reported to demonstrate the performance of the online RP in improving and maintaining the well-posedness of the model identification problem.

Table 3.1: Experimental conditions φ adopted for the collection of samples 1 to 3 in the experimental campaigns: inlet concentration of benzoic acid $C_{\mathrm{BA}}^{\mathrm{IN}}$; flowrate F; temperature of the oil bath T.

Sample number	$C_{\mathrm{BA}}^{\mathrm{IN}}$ $[\mathrm{mol\,L^{-1}}]$	F $[\mu\mathrm{L\,min^{-1}}]$	T [K]
1	1.50	20.0	413.0
2	1.00	10.0	393.0
3	1.25	15.0	403.0

3.4 Results

3.4.1 Simulated case: samples generated in-silico

Two campaigns of experiments, i.e., a non-RP campaign and a RP campaign, were simulated by integrating the kinetic model presented in Section 3.3.2 and adding Gaussian noise. Experimental conditions investigated in the course of the campaigns and the associated sampled concentrations of ethyl benzoate are given in Appendix A. The estimates for the kinetic parameters θ_1 and θ_2 for the non-RP campaign are reported in Table 3.2 together with information on their statistical quality. More specifically, the 95% confidence intervals and the correlation coefficient c_{12} between the kinetic parameters θ_1 and θ_2 are reported. One can see from Table 3.2 that the correlation coefficient c_{12} remains above 99.96% in the course of

Table 3.2: Simulated case: non-RP campaign. Parameter estimates are reported together with their respective 95% confidence intervals and correlation coefficient in the course of the experimental campaign. Parameter estimation and MBDoE problems are solved in the original parameter space Θ. The condition number of the log-likelihood function in Θ is reported in the table.

Simulated case - non-RP campaign

Samples collected	Estimates $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$ with 95% confidence intervals	Correlation coefficient c_{12}	p-value of target parameters θ^*	Condition number κ in Θ
1	[- , -]	-	-	-
2	[- , -]	-	-	-
3	[12.15 ± 2.14 , 6.56 ± 1.35]	0.9998	0.00%	$1.4{\cdot}10^4$
4	[14.83 ± 1.22 , 7.47 ± 0.81]	0.9996	0.00%	$6.1{\cdot}10^3$
5	[15.99 ± 1.01 , 7.85 ± 0.70]	0.9998	0.00%	$1.0{\cdot}10^4$
6	[15.06 ± 0.79 , 7.53 ± 0.53]	0.9997	0.00%	$7.2{\cdot}10^3$
7	[14.90 ± 0.74 , 7.47 ± 0.50]	0.9997	0.00%	$9.2{\cdot}10^3$
8	[14.84 ± 0.66 , 7.45 ± 0.44]	0.9997	0.00%	$8.2{\cdot}10^3$
9	[14.94 ± 0.63 , 7.49 ± 0.42]	0.9998	0.00%	$9.6{\cdot}10^3$

Table 3.3: Simulated case: RP campaign. Parameter estimates in the course of the experimental campaign are reported together with their respective 95% confidence intervals and correlation coefficient. Parameter estimation and MBDoE problems are solved in the transformed parameter space Ω. The condition number of the log-likelihood function in Ω is reported in the table.

Simulated case - RP campaign

Samples collected	Estimates $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$ with 95% confidence intervals	Correlation coefficient c_{12}	p-value of target parameters θ^*	Condition number κ in Ω
1	[- , -]	-	-	-
2	[- , -]	-	-	-
3	[16.44 ± 64.52 , 8.01 ± 25.05]	0.9999	0.00%	$5.5{\cdot}10^8$
4	[16.61 ± 3.55 , 8.06 ± 1.21]	0.9999	68.26%	$3.8{\cdot}10^2$
5	[15.60 ± 2.01 , 7.72 ± 0.68]	0.9998	86.41%	$1.2{\cdot}10^0$
6	[15.72 ± 1.62 , 7.76 ± 0.55]	0.9997	70.47%	$1.0{\cdot}10^0$
7	[15.72 ± 1.50 , 7.76 ± 0.51]	0.9998	69.84%	$1.0{\cdot}10^0$
8	[15.59 ± 1.44 , 7.71 ± 0.49]	0.9998	56.62%	$1.0{\cdot}10^0$
9	[15.39 ± 1.24 , 7.64 ± 0.42]	0.9998	64.74%	$1.0{\cdot}10^0$

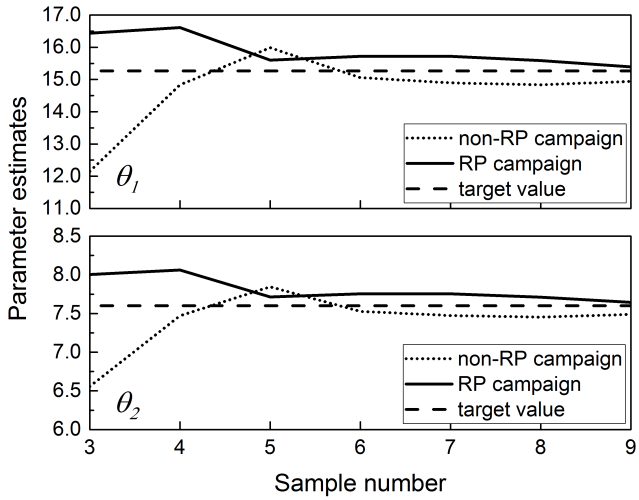


Figure 3.3: Simulated case: parameter estimates throughout the non-RP campaign (dotted) and the RP campaign (solid). The target parameter values are indicated by dashed lines.

the campaign. The parameter estimation and the MBDoE problems are solved in the original parameter space Θ where the condition number of the log-likelihood function remains above $6.1 \cdot 10^3$ throughout the whole experimental campaign. The χ^2-test was conducted to compare statistically the computed parameter distribution with the target parameter value θ^* (see Section 3.3.3 for information on how the test statistic is computed). As one can see from Table 3.2, a p-value of 0.00% in the course of the non-RP campaign suggests that the parameter estimates computed by the algorithm are statistically inconsistent with the target parameter values.

Parameter estimates and related information on their statistical quality are given in Table 3.3 for the RP campaign. In the course of the RP campaign, the correlation coefficient c_{12} remains above 99.97%. In the RP campaign, the parameter estimation problem and the MBDoE problem are solved in the transformed parameter space Ω, where the transformation matrix $\mathbf{G}$ is refined after the collection of each sample. The condition number of the log-likelihood function in Ω starts from a value of $5.5 \cdot 10^8$ at the first iteration of the model identification algorithm (i.e., after the collection of 3 samples) and it is reduced to 1.0 at the fourth iteration (i.e., after the collection of 6 samples). The benefit derived from the application of the online RP is validated by the χ^2-test. The p-value of the target value θ^* given the computed covariance at the end of the model identification campaign is 64.74%. This confirms that the algorithm computed estimates that are statistically consistent with the target parameter value θ^*.

The parameter estimates and related 95% confidence intervals obtained in the non-

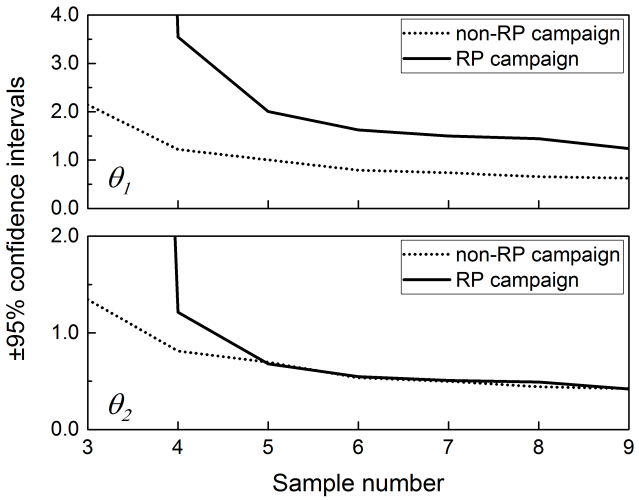


Figure 3.4: Simulated case: 95% confidence intervals associated with the parameter estimates throughout the non-RP campaign (dotted) and the RP campaign (solid).

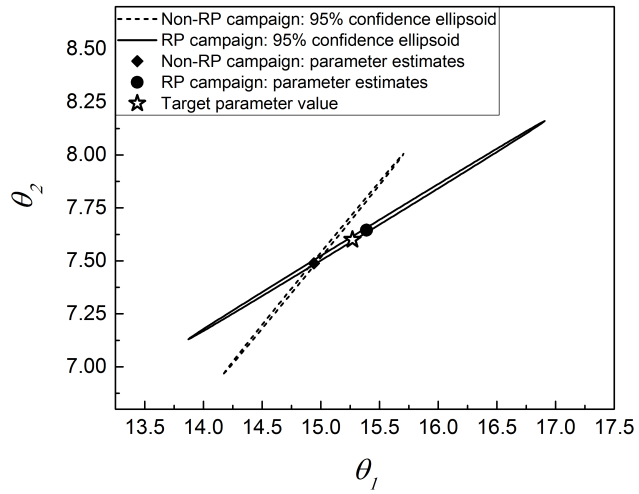


Figure 3.5: Simulated case: parameter estimates and related 95% confidence ellipsoids at the end of the non-RP campaign (dotted) and at the end of the RP campaign (solid). The target parameter value is highlighted in the graph by a star-shaped symbol.

RP campaign and in the RP campaign are compared graphically in Figure 3.3 and Figure 3.4. In Figure 3.3, one can see that both the methods present a similar convergence to the target parameter values, highlighted with dashed lines in the plot. In Figure 3.4, one can see that the 95% confidence intervals for the parameters are significantly different between the non-RP and the RP campaign. In particular the confidence interval of parameter $\hat{\theta}_1$ is significantly larger in the RP case than in the non-RP case. The discrepancy is interpreted as a consequence of an inaccurate computation of the log-likelihood gradient in the non-RP case, which results in an underestimation in the variance of the estimate $\hat{\theta}_1$.

The final estimates obtained in the non-RP and in the RP campaigns in the simulated

case are compared graphically in Figure 3.5. In Figure 3.5 the final parameter estimates are plotted with their respective 95% confidence ellipsoids for the non-RP campaign (dotted) and for the RP campaign (solid). The target parameter value is highlighted in Figure 3.5 by a star-shaped symbol. As one can see from Figure 3.5 the target value lies within the solid ellipsoid of the RP campaign, while it lies outside the dotted ellipsoid of the non-RP campaign. The graph shows that the non-RP campaign leads to the misleading conclusion that the target parameter values are not the parameters values of the physical system. The RP campaign led to a more robust estimate of the kinetic parameter values.

For both the non-RP and the RP campaign, derived estimates of pre-exponential factor and activation energy are available in Appendix A. Information on the goodness-of-fit after the collection of each sample is also reported in Appendix A.

Additional campaigns were performed in-silico to demonstrate that the performance of the model identification algorithm is insensitive to a change in the dataset, i.e., it is insensitive to a change in the random seed used to generate the data in-silico. The results obtained from 20 simulated campaigns are reported in Appendix B. Both in RP and non-RP campaigns, each algorithm iteration required only few seconds of CPU time.

3.4.2 Real case: samples collected from the experimental platform

Two campaigns of experiments, i.e., a non-RP campaign and a RP campaign, were performed on the automated system. Experimental conditions investigated in the course of the campaign and the associated sampled concentrations are given in Appendix C. Parameter

Table 3.4: Real case: non-RP campaign. Parameter estimates in the course of the experimental campaign with 95% confidence intervals and correlation coefficient. Parameter estimation and MBDoE problems are solved in the original parameter space Θ. The condition number of the log-likelihood function in Θ is reported in the table.

Real case - non-RP campaign

Samples collected	Estimates $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$ with 95% confidence intervals	Correlation coefficient c_{12}	Condition number κ in Θ
1	[- , -]	-	-
2	[- , -]	-	-
3	[16.16 ± 2.16 , 7.94 ± 1.49]	0.9998	$1.5{\cdot}10^4$
4	[16.44 ± 1.29 , 8.03 ± 0.89]	0.9996	$6.1{\cdot}10^3$
5	[17.15 ± 1.09 , 8.26 ± 0.77]	0.9998	$1.1{\cdot}10^4$
6	[16.80 ± 0.85 , 8.14 ± 0.59]	0.9997	$7.8{\cdot}10^3$
7	[17.23 ± 0.79 , 8.28 ± 0.56]	0.9998	$1.1{\cdot}10^4$
8	[17.15 ± 0.68 , 8.26 ± 0.48]	0.9997	$8.4{\cdot}10^3$
9	[17.42 ± 0.66 , 8.34 ± 0.47]	0.9998	$1.0{\cdot}10^4$

Table 3.5: Real case: RP campaign. Parameter estimates in the course of the experimental campaign with 95% confidence intervals and correlation coefficient. Parameter estimation and MBDoE problems are solved in the transformed parameter space Ω. The condition number of the log-likelihood function in Ω is reported in the table.

Real case - RP campaign

Samples collected	Estimates $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$ with 95% confidence intervals	Correlation coefficient c_{12}	Condition number κ in Ω
1	[- , -]	-	-
2	[- , -]	-	-
3	[17.54 ± 13.41 , 8.37 ± 5.38]	0.9999	$2.6{\cdot}10^{7}$
4	[18.12 ± 3.59 , 8.56 ± 1.23]	0.9999	$8.0{\cdot}10^{2}$
5	[16.86 ± 2.01 , 8.13 ± 0.68]	0.9998	$1.3{\cdot}10^{0}$
6	[16.90 ± 1.64 , 8.15 ± 0.55]	0.9997	$1.0{\cdot}10^{0}$
7	[16.91 ± 1.51 , 8.15 ± 0.51]	0.9998	$1.0{\cdot}10^{0}$
8	[16.83 ± 1.32 , 8.12 ± 0.45]	0.9997	$1.0{\cdot}10^{0}$
9	[16.98 ± 1.26 , 8.17 ± 0.43]	0.9998	$1.0{\cdot}10^{0}$

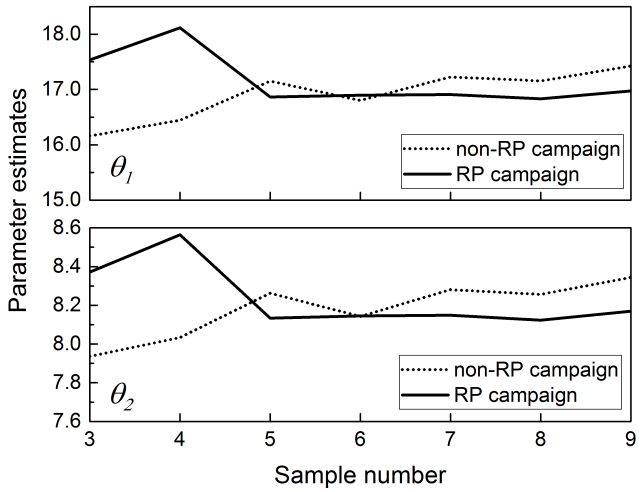


Figure 3.6: Real case: parameter estimates throughout the non-RP campaign (dotted) and the RP campaign (solid). The target parameter values are indicated by dashed lines.

estimates $\hat{\theta}$ with associated confidence intervals and correlation coefficient are reported in Table 3.4 for the non-RP campaign and in Table 3.5 for the RP campaign. Numerical estimates in terms of pre-exponential factor and activation energy were also computed from $\hat{\theta}$. These are reported in Appendix C.

In the course of the non-RP campaign (see Table 3.4), the parameter correlation c_{12} between $\hat{\theta}_1$ and $\hat{\theta}_2$ remains above 99.96%. In the non-RP campaign the parameter estimation and MBDoE problems are solved in the original parameter space Θ. The condition number of the log-likelihood function in Θ remains above $6.1 \cdot 10^3$ in the course of the non-RP campaign.

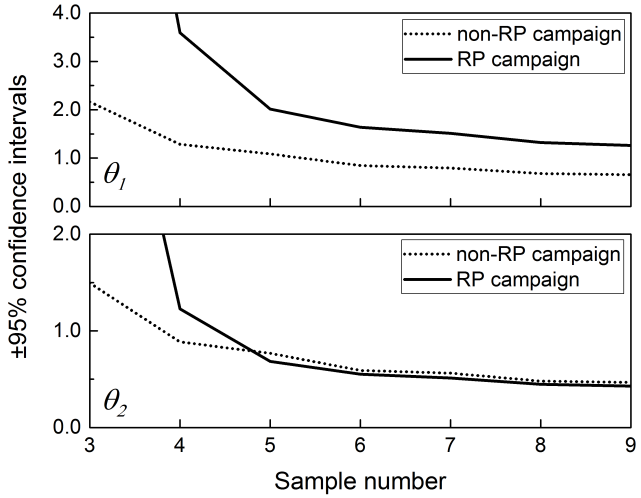


Figure 3.7: Real case: 95% confidence intervals associated with the parameter estimates throughout the non-RP campaign (dotted) and the RP campaign (solid).

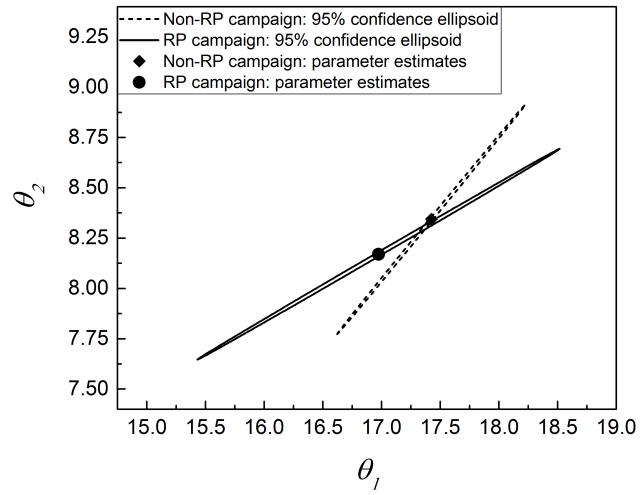


Figure 3.8: Real case: parameter estimates and related 95% confidence ellipsoids at the end of the non-RP campaign (dotted) and at the end of the RP campaign (solid).

The correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ is above 99.97% throughout the whole RP campaign (see Table 3.5). However, in the RP campaign, parameter estimation and MBDoE problems are solved in the transformed parameter space Ω. The condition number in Ω is reduced by the algorithm from an initial value of $2.6 \cdot 10^7$ to the minimum value 1.0 in four iterations (i.e., after the collection of 6 samples). The transformation matrix $\mathbf{G}$ is then adjusted after the collection of each sample to maintain a condition number $\kappa = 1.0$ until the end of the experimental campaign.

The parameter estimates and related 95% confidence intervals obtained in the non-RP and in the RP campaigns are plotted in Figure 3.6 and Figure 3.7. The 95% confidence

ellipsoids associated to the final parameter estimates achieved in the non-RP campaign and in the RP campaign are plotted in Figure 3.8.

Notice that in this case it is not possible to quantify and compare the performance of the two campaigns in retrieving the target parameter value. The target kinetic parameters are in fact unknown in the real case. One can observe from Figure 3.6 that the estimates achieved in the RP campaign exhibit a convergent behaviour around the values $\theta = [16.90, 8.15]^T$. Estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ in the non-RP campaign do not exhibit a convergent behaviour, but they tend to increase in the course of the non-RP campaign (see Figure 3.6). It is not possible to assess whether the absence of convergence in the non-RP campaign is the consequence of an unknown systematic disturbance in the system. However, it is possible to appreciate that the application of the online RP method led to the minimisation of the condition number (see Table 3.5) with the concomitant improvement in the numerical performance of the optimisation algorithms. Also in the real case, both in the RP and in the non-RP campaign, the CPU time required to complete each algorithm iteration was on the order of seconds.

A goodness-of-fit test was also conducted to demonstrate that the postulated first order single-reaction mechanism (see Section 3.3.2) provided an accurate representation of the chemical system. Nonetheless, it was recognised that an analysis on the goodness-of-fit was not significant for demonstrating the online RP method. It was chosen to report in Appendix C the numerical details regarding the analysis on the fitting quality.

3.4.3 Results discussion

Both in the simulated and in the real case, the 95% confidence intervals of the estimates after 9 collected samples differ significantly between the non-RP and the RP campaign (see Figure 3.4 and Figure 3.7). In the simulated case, a χ^2-test was conducted to compare the final statistics on the parameters computed in the RP campaign with the final statistics obtained in the non-RP campaign. It was shown that the confidence region of the parameter estimates computed in the RP campaign *contains* the target parameter value θ^* while the ellipsoid computed in the non-RP campaign does *not contain* the target value θ^* (see Figure 3.5). Hence, it was possible to demonstrate statistically that the campaign with online RP led to a more accurate quantification of the uncertainty region associated to the computed parameter estimates.

Figure 3.9a and Figure 3.9b show the condition numbers in the course of the non-RP and RP campaigns respectively. In the non-RP campaigns (see Figure 3.9a), the condition

number κ is around 10^4 and does not vary significantly in the course of the sample collection process. In the RP campaigns, both in the simulated and in the real case, the employment of the online RP method led to the minimisation of the condition number to $\kappa = 1.0$ in an initially ill-conditioned model identification problem (see Figure 3.9b). From Figure 3.9b, one can see that, both in the simulated and in the real case, the condition number is minimised to $\kappa = 1.0$ when sample 6 is collected, i.e., after 4 iterations in the model identification algorithm. This is explained by the fact that the update for the transformation matrix $\mathbf{G}$ is evaluated as a function of the Hessian $\mathbf{H}$ computed with the primary transformation matrix $\mathbf{G}_\mathrm{P}$ (see Section 3.2).

The condition number in the transformed space associated with $\mathbf{G}_\mathrm{P}$ may be very high at the first iteration of the algorithm. A high condition number at the *primary parameter estimation* step may lead to an inaccurate computation of the Hessian (i.e., an inaccurate quantification of the sloppiness) and, consequently, lead to the computation of an inappropriate update for $\mathbf{G}$. This does not appear to affect the performance of the online RP approach in the presented case study, but further analysis is required. It is object of future research activities to make the proposed algorithm insensitive towards numerical inaccuracies in the initial diagnosis of model sloppiness.

3.4.4 Computational times and problem size

The proposed model identification algorithm was applied online on the identification of a kinetic model involving ordinary differential equations. The model under study involves $N_\theta = 2$ model parameters. The numerical results presented in this Chapter were obtained

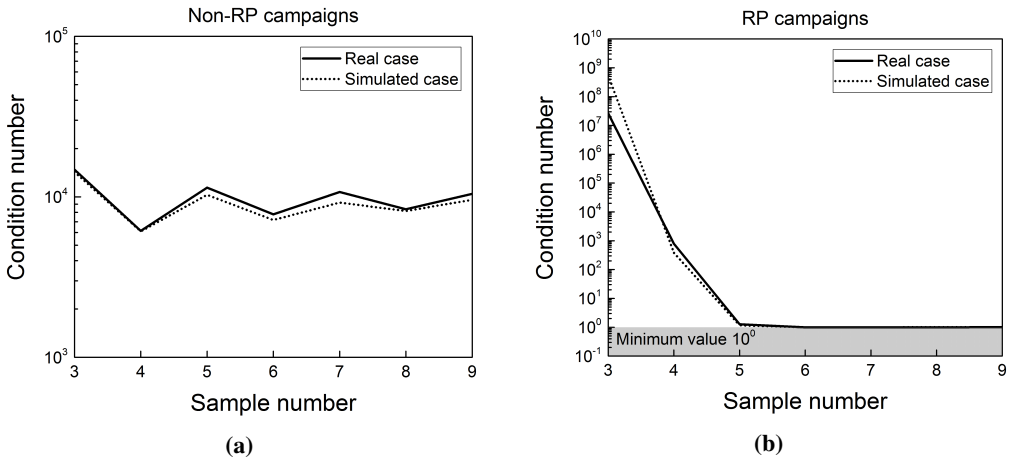


Figure 3.9: Condition number after each sample collected in the simulated case (dotted line) and in the real case (solid line): (a) non-RP campaigns; (b) RP campaigns.

on a 32-bit Windows machine with Intel® Core® i7-3770 3.40 GHz processor and 4.0 GB of RAM.

Table 3.6: Real case: Computational times associated with each algorithm iteration in the non-RP campaign and in the RP campaign.

Algorithm	Algorithm runtime [s]	
call	non-RP campaign	RP campaign
1	8.85	11.14
2	8.76	9.72
3	8.70	10.31
4	9.37	10.20
5	9.00	10.75
6	9.62	10.03
7	9.72	12.04

The computational runtime associated with each iteration of the model identification algorithm is reported in Table 3.6 for both the non-RP and the RP campaign performed on the automated experimental setup. In the course of the experimental campaigns, the model identification algorithm is called 7 times. As one can see from Table 3.6, the computational times associated with the algorithm calls in the RP campaign are higher than in the non-RP campaign. In the course of an algorithm call in the non-RP campaign, the *parametrisation update* stage and the *secondary parameter estimation* stage are not performed. The longest computation required in the RP case is associated primarily with the fact that the parameter estimation problem is solved two times in the course of an algorithm call. Nonetheless, the computational times in both campaigns are comparable for the considered case study, i.e., around 10.0 s per iteration both in the non-RP and in the RP case.

It is observed that in the presence of a higher number of parameters, the convergence rate of optimisation algorithms in the presence of a sloppy parametrisation may significantly decrease. Under such circumstance, an approach based on online RP may outperform a standard model identification algorithm also in terms of computational time. In fact, in the presence of a low condition number, a lower number of iterations and function evaluations is typically required to achieve convergence (Pyzara et al., 2011). Assessing the computational performance of the proposed approach in the presence of a higher number of parameters will be object of future research activities.

3.5 Final remarks

A parameter estimation algorithm implementing a novel approach of online reparametrisation, i.e., an approach of online transformation of the model parameter space, is proposed in this Chapter. The tool is designed specifically to reduce the chance of numerical failures associated with the estimation of parameters in the presence of *sloppy* model structures, i.e., models in which parameters are practically non-identifiable and/or extremely correlated.

The approach is based on two fundamental steps: *1*) a primary parameter estimation step, which is required to diagnose and quantify the sloppiness of the model parameter space; *2*) a parametrisation update step in which the sloppy parameter space is transformed into a robust space with the aim of reducing the sloppiness. Once the model parametrisation is updated, the parameter estimation is repeated solving an optimisation problem in the transformed, non-sloppy, parameter space. Additional samples are then designed by solving an optimal MBDoE problem in the transformed space with the aim of improving the statistical quality of the estimates. It is shown that numerical optimisation routines benefit significantly from the presence of a robust (i.e. non-sloppy) model parametrisation both at the parameter estimation and at the experimental design stage. Eventually, parameter estimates computed in the robust space are transformed to the original parameter space by applying algebraic transformations and returned as output to the user.

The performance of the presented algorithm was tested both in-silico and on a real system where an automated experimental platform was employed for online kinetic model identification. The objective in the case study was the estimation of the kinetic parameters in a two-parameter model of catalytic esterification of benzoic acid with ethanol in a flow reactor. Both in the simulated and in the real case, the algorithm iteratively reduced and eventually eliminated model sloppiness minimising the condition number of an originally ill-conditioned model identification problem. The minimisation of the condition number to unity and the concomitant elimination of model sloppiness resulted in an improved numerical robustness of the optimisation routines and matrix inversion functions employed in the course of the model identification process.

The proposed approach is particularly suited for implementation in autonomous model identification platforms. The reparametrisation method was integrated as an optional step in an online model identification algorithm implemented in a Python script. It was shown that the computational performance of the algorithm was not affected significantly by the

additional step of model reparametrisation. The modest computational cost associated with the reparametrisation step and the low memory requirement of the method makes it suitable for implementation also on embedded devices.

The numerical robustness of model identification algorithms towards model sloppiness represents a prerequisite for the application of the modelling frameworks illustrated in the following Chapters (see Figure 1.3). The next Chapter focuses on improving the robustness of model identification algorithms towards the presence of inappropriate modelling assumptions.

Chapter 4

Parameter estimation under structural model uncertainty

Part of this Chapter is adapted from the following articles:

Quaglio M., Fraga E. S., Cao E., Gavriilidis A., Galvanin F., A model-based data mining approach for determining the domain of validity of approximated models, *Chemometrics and Intelligent Laboratory Systems* 172, 2018, pp. 58-67

Quaglio M., Bezzo F., Gavriilidis A., Cao E., Al-Rifai N., Galvanin F., Identification of kinetic models of methanol oxidation on silver in the presence of uncertain catalyst behaviour, *AIChE Journal* 65(10), 2019, pp:e16707

Quaglio M., Fraga E. S., Galvanin F., Constrained model-based design of experiments for the identification of approximated models, *Proceedings of the 18th IFAC Symposium on System Identification* 2018, pp. 515-520

The author of this Thesis contributed to the above articles by developing the main novel ideas, implementing the simulations, and writing a significant part of the text. Hence, the author retains the right to include the articles in this Thesis since it is not published commercially and the journals are referenced as the original source.

4.1 Introduction

Parametric models derived from simplifying assumptions give an approximated description of the physical system under study. The practical applicability of an approximated model depends on the consciousness of its descriptive limits and on the *precise* estimation of its parameters. In this Chapter, a novel framework for the estimation of model parameters embracing structural model uncertainty is presented. In the framework, a model-based data

mining (MBDM) algorithm is used to estimate model parameters excluding the outliers from the fitting. A supervised machine learning classifier is then employed to detect patterns in the distribution of outliers to quantify the reliability of model predictions in unexplored regions of the experimental design space. The classifier returns a reliability map that can be used to constrain experimental design problems with the aim of collecting additional informative samples with low associated fitting cost.

4.2 Proposed methodology

An experimental setup is available to perform kinetic experiments on a physical system of interest. It is assumed that a preliminary dataset Y in the form (2.6) is collected for identifying a kinetic model for the process under study. The scientist proposes an approximated model structure in the form

$$\begin{aligned} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t, \theta) &= \mathbf{0} \\ \hat{\mathbf{y}} &= \mathbf{h}(\mathbf{x}, \mathbf{u}, t, \theta) \end{aligned} \tag{2.1}$$

It is assumed that the objective of the scientist is to complete the identification of the model, which requires both *i*) the precise estimation of the parameters θ by fitting experimental data and *ii*) the computation of the domain of the model reliability, namely the range of conditions in which the identified model is expected to provide accurate predictions. A framework for the identification of approximated models is proposed in Figure 4.1 to address the multi-objective task of both parameter estimation and the determination of the domain of model reliability given the available experimental evidence. The approach starts from the availability of a preliminary set of experimental data Y and an approximated model structure in the form (2.1). The procedure involves three fundamental steps:

1. *A Model-based data mining step*. The model parameters are fitted to the available dataset Y employing a tailored approach for robust regression (Rousseeuw and Leroy, 1987), namely a model-based data mining (MBDM) method for parameter estimation. MBDM produces two outputs: *i*) it classifies the observed experimental conditions φ_i (with $i = 1,...,N$) as *compatible* or *incompatible* with the candidate model following a criterion based on the the quality of fitting and *ii*) it computes the maximum likelihood estimate for the parameters fitting only the model compatible data.

2. *A Support Vector Machine training step*. The classified conditions φ_i with $i = 1,...,N$, returned by MBDM at step 1, are used to train a Support Vector Machine (SVM) clas-

sifier (Schölkopf and Smola, 2002; Smola and Schölkopf, 2004), which generalises the classification to unexplored experimental conditions. SVM returns a model reliability map $I(\varphi)$ which quantifies the expected model accuracy across the experimental design space.

3. *A constrained MBDoE step*. If the parameter estimates computed by MBDM at step 1 do not meet the desired statistical requirements, then additional informative samples should be collected and included in the parameter estimation problem. The following experiments are designed employing known MBDoE criteria (see Section 2.4.6). The optimal experimental design problem is bounded within the model reliability domain (i.e. the design is constrained to conditions φ such that $I(\varphi) > 0$) to prevent the collection of further model incompatible data.

In the following sections, the aforementioned steps are further detailed. The MBDM approach is illustrated in Section 4.2.1. The underlying mathematics of SVM is then presented in Section 4.2.2. The constrained MBDoE problem is formulated in Section 4.2.3.

4.2.1 Model-Based Data Mining for Parameter Estimation

If the structure (2.1) is approximated, one shall not expect the model to be accurate across the entire experimental design space. Rousseeuw and Leroy (1987) and Buzzi-Ferraris and Manenti (2009) recognised that data collected at conditions where the modelling assumptions are not valid are not significant for the estimation of the model parameters and shall be regarded as outliers (for more information on outlier types see Section 2.6). In fact, equivalently to outliers caused by gross measurement error and/or system disturbances, the fitting of these data may lead to estimates with debatable physical significance and the identification of a model with poor predictive performance. Nevertheless, the domain in which the approximated modelling assumptions are valid is normally not known a-priori.

A heuristic model-based data mining (MBDM) approach is proposed to detect the presence of outliers in the dataset and estimate the model parameters neglecting model-incompatible data. The robust weighted least square estimator proposed by Rousseeuw and Leroy (1987) is employed as an MBDM tool. The estimator was introduced in Section 2.6. MBDM requires the solution of the optimisation problem in (2.27) where the function $\mathscr{L}_{\mathrm{DM}}$

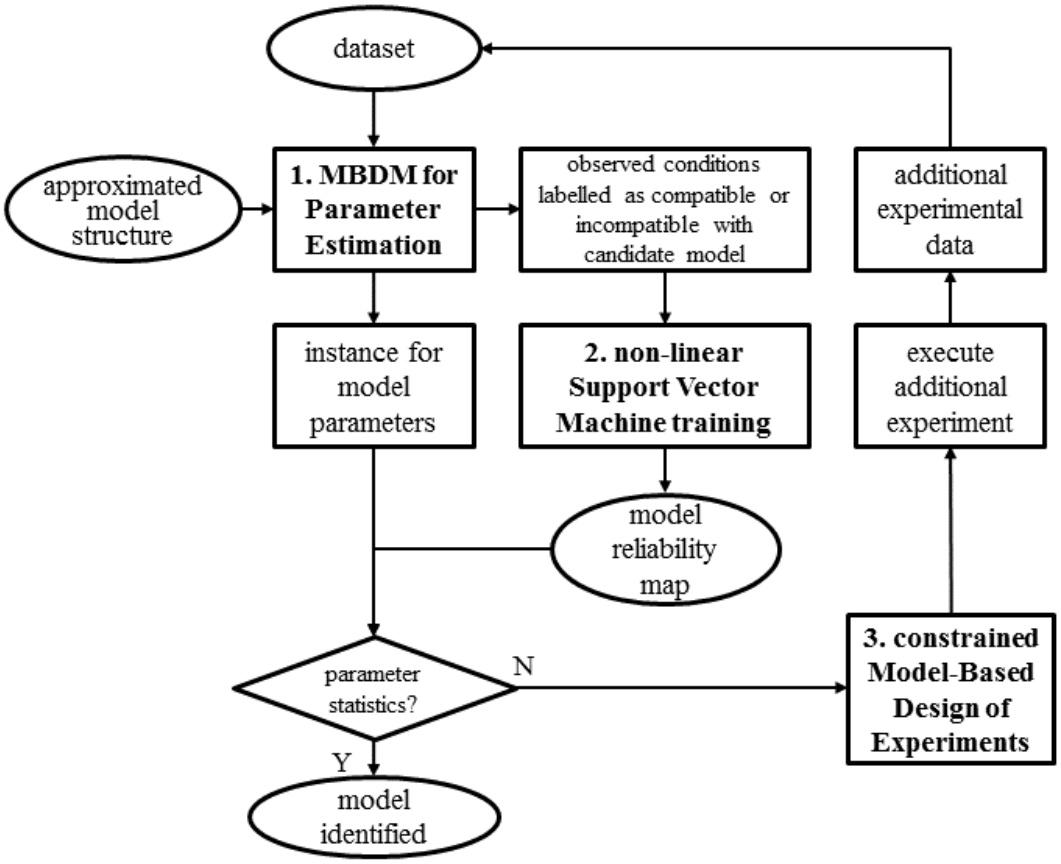


Figure 4.1: Proposed framework for model identification. Boldface blocks represent fundamental steps in the proposed methodology.

to maximise is given in (2.28).

$$\hat{\theta}_{DM} = \arg\max_{\theta\in\Theta} \mathscr{L}_{DM} \tag{2.27}$$

$$\mathscr{L}_{DM} = \sum_{i=1}^{N} \frac{1+\beta_i}{2} \cdot \{N_y c^2 - [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]\} \tag{2.28}$$

$$\text{s.t.} \quad \beta_i(\theta) = \begin{cases} +1 & \text{if} \quad [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)] \leq N_y c^2 \\ -1 & \text{if} \quad [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta)] > N_y c^2 \end{cases} \quad \forall \quad i \tag{2.29}$$

In (2.28), the quantities $\beta_i \in \{+1, -1\}$ with $i = 1, ..., N$ represent binary weights, which are introduced to control the inclusion ($\beta_i = +1$) or exclusion ($\beta_i = -1$) of samples in the objective function $\mathscr{L}_{DM}$. The conditions in (2.29) ensure that samples are considered for

parameter fitting only if the associated residuals are small. The quantity c quantifies the maximum threshold of acceptance for a model residual.

Given a reasonable choice of the hyperparameter c, the solution of the MBDM problem in (2.27) leads to the automated exclusion from the parameter estimation problem of the samples that are statistically incompatible with the modelling assumptions. The quantity $\hat{\theta}_{DM}$ represents a robust maximum likelihood estimate obtained from the fitting of the possibly reduced dataset Y' as in (2.30). The covariance associated with the estimate $\hat{\theta}_{DM}$ is calculated as

$$\mathbf{V}_\theta = [-\nabla\nabla^T \mathscr{L}(Y'|\hat{\theta}_{DM})]^{-1} \tag{2.31}$$

Detected outliers may be classified either as *1*) samples collected outside the domain of model reliability *2*) samples collected in the presence of significant systematic errors or *3*) samples collected in the presence of significant system disturbances. Notice that MBDM does not distinguish between these three outlier classes. A possible practical way to classify the outliers is to repeat the sampling. If the incompatibility persists after the repetition, the outlier shall be classified in the first or in the second category, i.e. the sample is collected outside the domain of model reliability or in the presence of systematic errors. If the repeated sample is instead found to be compatible, the incompatibility detected before the repetition shall be interpreted as the consequence of a system disturbance.

4.2.2 Support Vector Machine training

The solution of the MBDM problem in (2.27), leads to the construction of a function $\varphi_i \rightarrow \hat{\beta}_i \in \{1,-1\}$ with $i = 1,...,N$ (where $\hat{\beta}_i = \beta_i(\hat{\theta}_{\mathrm{DM}})$), which classifies the observed experimental conditions φ_i, with $i = 1,...,N$, either as compatible or incompatible with the candidate model. It is now of interest to identify a decision function $I(\varphi)$, based on the training set $\{(\varphi_i, \hat{\beta}_i)|i = 1,...,N\}$, whose sign can be used to classify the performance of the model in unexplored experimental conditions. A decision function is required to quantify *i*) the reliability on the model predictions across the space of experimental conditions and *ii*) the expected model fitting quality across the experimental design space for supporting the design of new trials to enhance parameter precision.

In the proposed approach, the classification of the observed experimental conditions is generalised to a generic set of conditions $\varphi \in \Phi$ employing a non-linear Support Vector Machine classifier (Cortes and Vapnik, 1995) with Gaussian kernel K (Schölkopf and Smola,

2002). The Gaussian kernel, also known as the radial basis function, is defined as

$$K(\varphi_i, \varphi_j) = e^{-\frac{(\varphi_i - \varphi_j)^T (\varphi_i - \varphi_j)}{2\gamma^2}} \tag{4.1}$$

where the hyperparameter γ represents a decay length, which quantifies the degree of *similarity* between two different sets of experimental conditions φ_i and φ_j. The Gaussian kernel is employed for its generality of application and for its capability of computing decision functions with non-linear, non-connected and non-convex geometry.

The application of the non-linear SVM classifier results in the construction of a decision function $I(\varphi)$, namely a model reliability map, in the form (4.2) whose sign is used to classify unexplored conditions of the experimental design space in terms of acceptable ($I > 0$) or unacceptable ($I < 0$) expected model performance.

$$I(\varphi) = \sum_{i=1}^{N} \hat{\alpha}_i \hat{\beta}_j K(\varphi, \varphi_i) + b \tag{4.2}$$

In (4.2), b represents the offset of the decision function and $\hat{\alpha}_i$ with $i = 1, ..., N$ are the values for the Lagrange multipliers obtained through the solution of the following convex optimisation problem (Cortes and Vapnik, 1995):

$$\begin{aligned} \max_{\alpha_1, ..., \alpha_N} & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \hat{\beta}_i \hat{\beta}_j K(\varphi_i, \varphi_j) \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i \hat{\beta}_i = 0, \\ & 0 \leq \alpha_i \leq C_i \quad \forall i = 1, ..., N \end{aligned} \tag{4.3}$$

The value for the parameter b in (4.2) is computed after the solution of the optimisation problem (4.3) from the *Karush-Kuhn-Tucker* complementarity condition associated to any margin support vector (characterised by the condition $\alpha_i > 0$) (Burges, 1998). In (4.3), C_i (with $i = 1, ..., N$) are regularisation parameters that can be adjusted to modify the weight of each sample in the decision function.

Since SVMs are sensitive to the scale of the input space, experimental conditions are normalised before the application of the learning machine. Notice that a number of degrees of freedom are present in the problem due to the regularisation parameters C_i and the decay length of the radial basis function γ. The hyperparameters C_i and γ may be chosen *a priori* or following heuristic rules (King and Zeng, 2001). Alternatively, in the presence

of a sufficiently large dataset, an optimal hyperparameter set may be identified through cross-validation (Bergstra and Bengio, 2012).

4.2.3 Constrained Model-Based Design of Experiments

From the covariance $\mathbf{V}_\theta$ computed according to (2.31) it is possible to assess the statistical quality of the parameter estimates, e.g. by performing a t-test. In case of unsatisfactory parameter statistics, additional samples should be collected from the setup and included in the parameter estimation problem. It is assumed that the scientist is willing to design N_{sp} additional samples with the aim of reducing parameter uncertainty employing an MBDoE approach (see Section 2.4.6).

MBDoE is formulated as an optimisation problems in which the function to be minimised is a measure ψ (e.g. the trace or the determinant) of the predicted covariance matrix $\hat{\mathbf{V}}_\theta$, which is calculated as in (2.21). However, conventional MBDoE methods for parameter precision do not consider the presence of structural model uncertainty in the formulation of design metrics based on Fisher information. Hence, in the presence of an approximated model structure, MBDoE methods may lead to the design of experiments in conditions $\varphi \in \Phi$ where the model is particularly inaccurate. Samples collected outside the domain of model reliability may be rich in terms of Fisher information, but their fitting could result in an unacceptable degradation of the model fitting quality and a loss of model predictive performance. In this work, a conservative approach to MBDoE is proposed where the experimental design is *constrained* within the domain of model reliability, i.e., at conditions $\varphi \in \Phi | I(\varphi) \geq 0$ in which the model is expected to provide a good fitting.

$$\begin{aligned} \varphi_1^*, ..., \varphi_{N_{sp}}^* = \arg \min_{\varphi_1,...,\varphi_{N_{sp}}} \psi(\varphi_1, ..., \varphi_{N_{sp}})|_{\theta=\hat{\theta}_{DM}} \\ \text{s.t.} \quad \varphi_k \in \Phi | I(\varphi_k) \geq 0 \quad \forall k = 1, ..., N_{sp} \end{aligned} \tag{4.4}$$

The constrained MBDoE problem is formulated in (4.4), where $\varphi_1^*, ..., \varphi_{N_{sp}}^*$ represent the optimised experimental conditions for the collection of the additional samples. A sketch to illustrate the procedure is proposed in Figure 4.2. The top-left colourmap in Figure 4.2 represents the distribution of the design metric $\psi(\varphi)$ across the experimental design space. The bottom-left graph in Figure 4.2 shows the reliability map $I(\varphi)$. In the constrained MBDoE problem, the information metric is maximised within the model reliability domain, i.e. at conditions $I > 0$, as shown in the right graph in Figure 4.2.

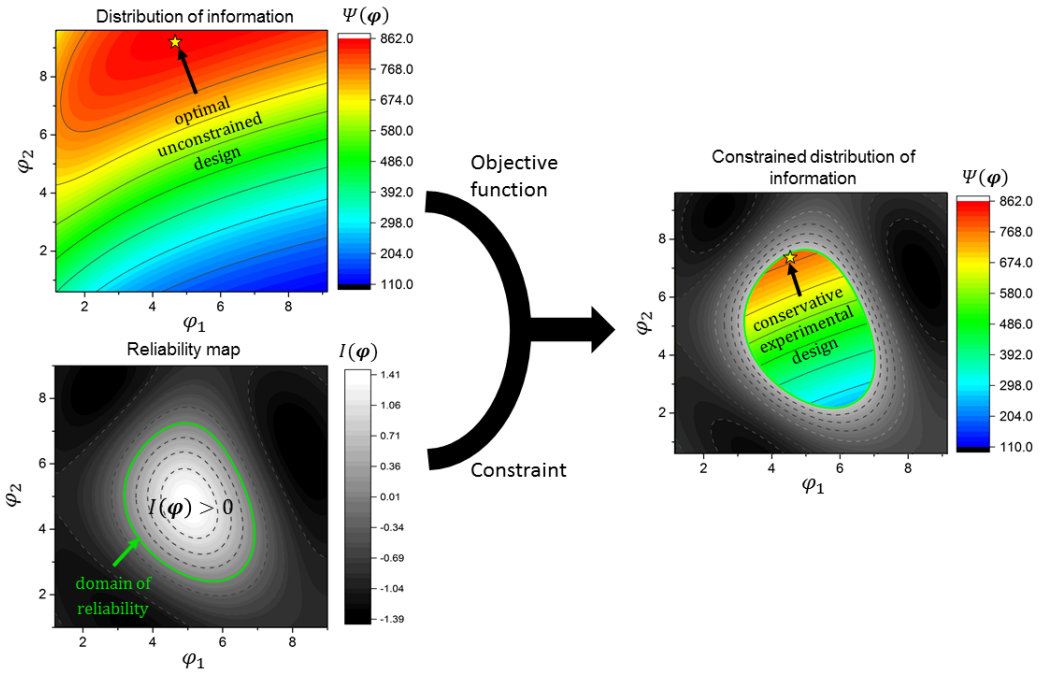


Figure 4.2: Qualitative sketch illustrating the proposed procedure for the identification of optimal informative experimental conditions within the model reliability domain.

4.3 Case studies

The model identification approach presented in Section 4.2 is tested on a simulated case study in Section 4.3.1 where the aim is the online identification of an approximated kinetic model of ethanol dehydrogenation on a copper/copper-chromite based catalyst. The case study is inspired by the work of Carotenuto and co-workers (Carotenuto et al., 2013). The approximated model considers two reactions and its identification requires the estimation of $N_\theta = 4$ kinetic parameters. A second case study is proposed in Section 4.3.2, where the aim is the identification of an approximated kinetic model of methanol oxidation on silver catalyst (Andreasen et al., 2005) in a continuous flow microreactor (Galvanin et al., 2015). In this case, the approximated model involves three reactions and its identification requires the estimation of $N_\theta = 2$ kinetic constants. In this case study the identification of the model is performed offline using real experimental data.

4.3.1 Case study 1: ethanol dehydrogenation on copper

4.3.1.1 System model

The catalytic reaction of ethanol dehydrogenation is assumed to occur in an ideal packed-bed tubular reactor. It is assumed that the reaction occurs at isothermal conditions in the

absence of pressure drops and mass transfer limitations. The space evolution of the reacting gaseous mixture is described by the set of differential equations (4.5). Five chemical species are considered, i.e.: ethanol CH_3CH_2OH (EtOH); acetaldehyde CH_3CHO (AcH); ethyl acetate $CH_3COOCH_2CH_3$ (EA); hydrogen H_2; and nitrogen N_2 (used as inert carrier).

$$\frac{d\dot{n}_i(z)}{dz} = w \sum_{j=1}^{N_R} \nu_{ij} r_j \quad \forall\, i = \mathrm{EtOH}, \mathrm{AcH}, \mathrm{EA}, \mathrm{H_2}, \mathrm{N_2} \tag{4.5}$$

In (4.5), z is the axial coordinate of the tubular reactor normalised on the catalyst bed length; $\dot{n}_i$ $[\mathrm{mol\,h^{-1}}]$ is the molar flowrate of the i-th component of the mixture; w [g] is the catalyst weight; N_R is the number of reactions; ν_{ij} is the stoichiometric coefficient of the i-th species in the j-th reaction; r_j $[\mathrm{mol\,h^{-1}g^{-1}}]$ is the reaction rate of the j-th reaction normalised on the catalyst weight.

In this study, the Langmuir-Hinshelwood-Hougen-Watson (LHHW) kinetics proposed by Carotenuto *et al.* is adopted as the *true* model of the physical system (Carotenuto et al., 2013). The kinetic model involves $N_R = 3$ reactions whose stoichiometry is

$$\begin{aligned}
&\text{Reaction 1:} && \text{Ethanol} \leftrightharpoons \text{Acetaldehyde} + \mathrm{H_2} \\
&\text{Reaction 2:} && \text{Ethanol} + \text{Acetaldehyde} \leftrightharpoons \text{Ethyl Acetate} + \mathrm{H_2} \\
&\text{Reaction 3:} && \text{Acetaldehyde} \rightarrow \text{other products}
\end{aligned} \tag{4.6}$$

Reaction 1 describes the step of ethanol dehydrogenation into acetaldehyde, reaction 2 accounts for the formation of ethyl acetate from ethanol and acetaldehyde and reaction 3 accounts for parallel reactions consuming acetaldehyde to give side undesired products. The reaction rates are

$$\begin{aligned}
r_1 &= \frac{A_1 e^{-\frac{E_{a1}}{RT}} b_{\mathrm{EtOH}} P_{\mathrm{EtOH}} \left(1 - \left(1/K_{eq1}\right)\left(\frac{P_{\mathrm{AcH}} P_{\mathrm{H_2}}}{P_{\mathrm{EtOH}}}\right)\right)}{\left(1 + b_{\mathrm{EtOH}} P_{\mathrm{EtOH}} + b_{\mathrm{AcH}} P_{\mathrm{AcH}} + b_{\mathrm{EA}} P_{\mathrm{EA}} + b_{\mathrm{H_2}} P_{\mathrm{H_2}}\right)^2} \\
r_2 &= \frac{A_2 e^{-\frac{E_{a2}}{RT}} b_{\mathrm{EtOH}} b_{\mathrm{AcH}} P_{\mathrm{EtOH}} P_{\mathrm{AcH}} \left(1 - \left(1/K_{eq2}\right)\left(\frac{P_{\mathrm{EA}} P_{\mathrm{H_2}}}{P_{\mathrm{EtOH}} P_{\mathrm{AcH}}}\right)\right)}{\left(1 + b_{\mathrm{EtOH}} P_{\mathrm{EtOH}} + b_{\mathrm{AcH}} P_{\mathrm{AcH}} + b_{\mathrm{EA}} P_{\mathrm{EA}} + b_{\mathrm{H_2}} P_{\mathrm{H_2}}\right)^2} \\
r_3 &= A_3 e^{-\frac{E_{a3}}{RT}} P_{\mathrm{AcH}}^2
\end{aligned} \tag{4.7}$$

where A_j $[\mathrm{mol\,g^{-1}h^{-1}}]$ and E_{aj} $[\mathrm{J\,mol^{-1}}]$ are respectively the pre-exponential factor and the activation energy of the j-th reaction; R is the ideal gas constant $[\mathrm{J\,mol^{-1}K^{-1}}]$ and T is temperature [K]. Parameter b_i $[\mathrm{bar^{-1}}]$ is the adsorption coefficient related to the i-th mixture

component. P_i [bar] is the partial pressures of the i-th chemical species and it is defined as $P_i = (\dot{n}_i / \sum_i \dot{n}_i) P_{TOT}$, where P_{TOT} [bar] is the total pressure in the gas bulk. Quantities K_{eq1} and K_{eq2} are the equilibrium constants for reaction 1 and reaction 2 respectively. The equilibrium constants are evaluated from the Van't Hoff equation (Carotenuto et al., 2013) as illustrated in Appendix D. The values of the kinetic parameters estimated by Carotenuto et al. (2013) are assumed as the *true* kinetic parameters of the system. Kinetic parameter values associated with the LHHW model are reported in Table D.1 in Appendix.

4.3.1.2 Approximated model

The identification of the LHHW system model described in Section (4.3.1.1) requires the estimation of 10 kinetic parameters, i.e. A_1, E_{a1}, A_2, E_{a2}, A_3, E_{a3}, b_{EtOH}, b_{AcH}, b_{EA}, $b_{\mathrm{H_2}}$. It is assumed that the amount of resources to perform the experiments is insufficient for the identification of a comprehensive LHHW model and a compromise between model complexity and model accuracy is preferred. The scientist proposes an approximated kinetic model which involves only reaction 1 and reaction 2 of the total mechanism (4.6). Furthermore, the scientist also suggests to model the rates for reaction 1 and 2 as simple power laws. The approximated reaction rates are

$$\begin{aligned} r_1 &= A_1 e^{-\frac{E_{a1}}{RT}} P_{\mathrm{EtOH}} \Big(1 - \big(1/K_{eq1}\big)\Big(\frac{P_{\mathrm{AcH}} P_{\mathrm{H_2}}}{P_{\mathrm{EtOH}}}\Big)\Big) \\ r_2 &= A_2 e^{-\frac{E_{a2}}{RT}} P_{\mathrm{EtOH}} P_{\mathrm{AcH}} \Big(1 - \big(1/K_{eq2}\big)\Big(\frac{P_{\mathrm{EA}} P_{\mathrm{H_2}}}{P_{\mathrm{EtOH}} P_{\mathrm{AcH}}}\Big)\Big) \\ r_3 &= 0 \end{aligned} \tag{4.8}$$

The approximated kinetic model in (4.8) only involves four kinetic parameters, i.e. $\theta = [A_1, E_{a1}, A_2, E_{a2}]$.

4.3.1.3 Objective and methods

The identification of the approximated kinetic model (4.8) requires the precise estimation of the kinetic parameters $\theta = [A_1, E_{a1}, A_2, E_{a2}]$ and the determination of the model reliability domain. A positive t-test with 95% of significance is set as statistical requirement for the parameters. The following assumptions are made:

1. *Design space.* A three dimensional experimental design space is assumed $\Phi = (\dot{n}_{\mathrm{EtOH}}|_{z=0}, P_{TOT}, T)$ where the manipulable experimental conditions are: ethanol molar inlet flowrate $\dot{n}_{\mathrm{EtOH}}|_{z=0}$ (range $0.1 - 2.5$ mol h^{-1}); the total pressure P_{TOT} (range $10 - 30$ bar); temperature T (range 453-533 K). The inlet molar flowrate of the other

species is fixed at $[\dot{n}_{\mathrm{AcH}}, \dot{n}_{\mathrm{EA}}, \dot{n}_{\mathrm{H_2}}, \dot{n}_{\mathrm{N_2}}]|_{z=0} = [0.0, 0.0, 0.057, 0.057]$ mol h^{-1}. The catalyst weight is fixed at $w = 2.0$ g.

2. *Measurements and errors.* It is assumed that the molar flowrates of ethanol, acetaldehyde, ethyl acetate and hydrogen at the outlet are the measurable output variables in the system. Measurements are generated employing the system model illustrated in Section 4.3.1.1) adding uncorrelated Gaussian noise with covariance $\Sigma_y = 2.25 \cdot 10^{-4}\mathbf{I}$ mol^2 h^{-2}.

3. *Preliminary dataset.* A preliminary dataset with $N = 8$ sample is available to compute a preliminary estimate for the model parameters. The preliminary dataset is obtained from the simulation of a full factorial design with three factors (i.e. ethanol inlet flowrate, total pressure and temperature) and two levels for each factor.

Two cases are presented and compared:

ML case The model parameters are estimated with a conventional Maximum Likelihood approach (Bard, 1974) and additional samples are designed with standard MBDoE methodologies for parameter precision (Pukelsheim, 2006). Parameter estimates $\hat{\theta}$ are updated after the collection of every sample. A sequential D-optimal MBDoE is employed for designing the samples. The number of samples N_{sp} that are simultaneously designed at every iteration is chosen iteratively in the range $N_{sp} = 1, ..., N_{sp}^{MAX}$ (where N_{sp}^{MAX} is set equal to 3) to evaluate the minimum number of experiments required to meet the desired parameter statistics. Once N_{sp} experiments are designed, the algorithm selects and performs the k-th most informative designed sample according to $k = \arg\max_{k=1,...,N_{sp}} \mathrm{Tr}(\hat{\mathbf{H}}_k)$. The campaign stops when all parameters pass the 95% t-test or when the maximum number of samples N^{MAX} is collected.

MBDM case The model is identified employing the methodology presented in Section 4.2. Model parameters are estimated employing a MBDM estimator and additional samples are designed with a constrained MBDoE approach. The following settings are adopted in the MBDM case:

1. *MBDM settings.* The MBDM problem is formulated as in (2.27) imposing a maximum discrepancy tolerance $c = 2.0$. This is equivalent to treating as outliers the normalised residuals that exceed the range of 2 standard deviations of measurement noise.

2. *SVM settings.* The experimental design space is normalised to the unit cube before the training of SVM. The SVM classifier implemented in the Python package *scikit-learn* (Pedregosa et al., 2011) is used. The hyperparameters of the learning machine are set *a priori*: the decay length γ of the radial basis function is set to its default value $\gamma = 1.0$ in *scikit-learn*; C_j are computed from the *balanced* class_weight module of *scikit-learn* (King and Zeng, 2001) to account for the possibly very different number of compatible and incompatible samples in the dataset.
3. *Constrained-MBDoE settings.* Additional samples are designed following the same criteria as in the ML case, but bounding the MBDoE problem within the domain of model reliability computed by SVM.

Estimates $\hat{\theta}_{DM}$ and reliability map $I(\varphi)$ are updated after the collection of every sample. The procedure stops once all parameters pass the 95% t-test or when the maximum number of allowed samples N^{MAX} is reached.

A script to conduct the case study was implemented in Python 2.7. The solver SLSQP implemented in the *scipy* package (Jones et al., 2001) is employed at every iteration of the procedure for both the parameter estimation and the experimental design steps.

4.3.1.4 Results and discussion

The parameter estimates $\hat{\theta}$ and the associated sum of squared residuals χ^2_Y in the ML case are reported in Table 4.1. When the parameters are estimated by using a standard maximum likelihood approach, all the available samples are fitted. The sum of squared residuals is $\chi^2_Y = 88.41$ and the model is falsified for under-fitting by the 90% goodness-of-fit test. The information content of the full factorial preliminary design is sufficient for estimating precisely all the model parameters, i.e. all the estimates pass the 95% t-test and there is no necessity to collect additional samples.

Parameter estimates and model statistics in the course of the simulated campaign in the MBDM case are reported in Table 4.2. One can see that the desired parameter statistics are achieved after the collection of 8 additional samples, i.e. the identification of the model required the collection of 16 samples in total. In the course of the constrained experimental campaign, the reliability map $I(\varphi)$ is updated based on all the observed experimental conditions and the labelling computed by MBDM. The dynamic behaviour of the reliability function can be appreciated in the plots of Figure 4.3, where the reliability boundary,

Table 4.1: ML case: Parameter estimates and model statistics in the course of the experimental campaign.

Collected samples N	Fitted samples	Parameter estimates* $\hat{\theta} = [A_1\ E_{a1}\ A_2\ E_{a2}]$					χ^2_Y**	$\chi^2(95\%)$
8	8	[	$2.62 \cdot 10^{-1}$	$3.96 \cdot 10^{1}$	$1.63 \cdot 10^{-3}$	$1.42 \cdot 10^{1}$]	**88.41	36.42

*the parameter did not pass the t-test with 95% of significance
**A χ^2_Y larger than $\chi^2(95\%)$ indicates that the 90% goodness-of-fit test is failed for under-fitting

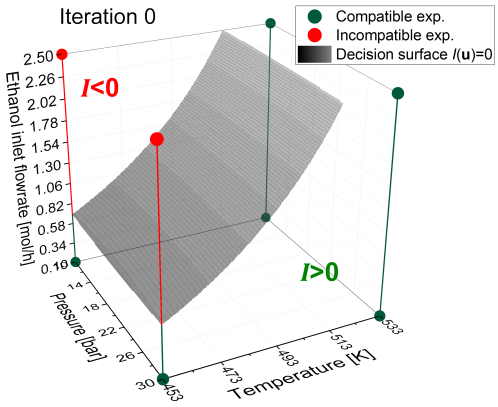
Table 4.2: MBDM case: Parameter estimates and model statistics in the course of the experimental campaign.

Collected samples N	Fitted samples	Parameter estimates* $\hat{\theta}_{\mathrm{DM}} = [A_1\ E_{a1}\ A_2\ E_{a2}]$					χ^2_Y**	$\chi^2(95\%)$
8	6	[	$*3.56 \cdot 10^{-1}$	$*3.95 \cdot 10^{1}$	$*1.99 \cdot 10^{-3}$	$*1.41 \cdot 10^{1}$]	21.10	36.42
9	7	[	$*4.20 \cdot 10^{-1}$	$*3.92 \cdot 10^{1}$	$*2.18 \cdot 10^{-3}$	$*1.40 \cdot 10^{1}$]	25.84	41.34
10	8	[	$*4.17 \cdot 10^{-1}$	$*3.91 \cdot 10^{1}$	$*2.24 \cdot 10^{-3}$	$*1.41 \cdot 10^{1}$]	29.20	46.20
11	9	[	$*3.13 \cdot 10^{-1}$	$3.91 \cdot 10^{1}$	$*1.78 \cdot 10^{-3}$	$*1.40 \cdot 10^{1}$]	32.52	51.00
12	8	[	$*6.03 \cdot 10^{-1}$	$*3.89 \cdot 10^{1}$	$*3.62 \cdot 10^{-3}$	$*1.37 \cdot 10^{1}$]	31.68	46.20
13	8	[	$7.67 \cdot 10^{-1}$	$3.84 \cdot 10^{1}$	$*4.46 \cdot 10^{-3}$	$*1.36 \cdot 10^{1}$]	31.84	46.20
14	10	[	$*2.79 \cdot 10^{-1}$	$3.93 \cdot 10^{1}$	$1.71 \cdot 10^{-3}$	$1.32 \cdot 10^{1}$]	48.69	55.76
15	10	[	$*4.41 \cdot 10^{-1}$	$3.64 \cdot 10^{1}$	$2.00 \cdot 10^{-3}$	$1.35 \cdot 10^{1}$]	36.93	55.76
16	11	[	$4.41 \cdot 10^{-1}$	$3.64 \cdot 10^{1}$	$1.97 \cdot 10^{-3}$	$1.35 \cdot 10^{1}$]	45.57	60.84

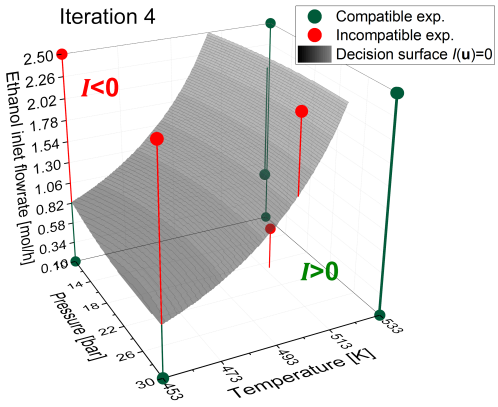
*the parameter did not pass the t-test with 95% of significance
**A χ^2_Y larger than $\chi^2(95\%)$ indicates that the 90% goodness-of-fit test is failed for under-fitting

defined by $I(\varphi) = 0$, is reported after the collection of the preliminary 8 samples (Figure 4.3a), after 12 collected samples (Figure 4.3b) and after 16 collected samples (Figure 4.3c). The dots in the plots of Figure 4.3 represent the conditions associated with the collected samples and the colour indicates the labelling computed by MBDM at the given iteration: *compatible* samples ($\hat{\beta}_i = +1$) are marked with green dots; *incompatible* samples ($\hat{\beta}_i = -1$) are marked with red dots. Experimental conditions and generated samples in the course of the experimental campaign in the MBDM case are reported in Table D.2 in Appendix. In Table D.2, the final values of the labels computed by MBDM are also given.

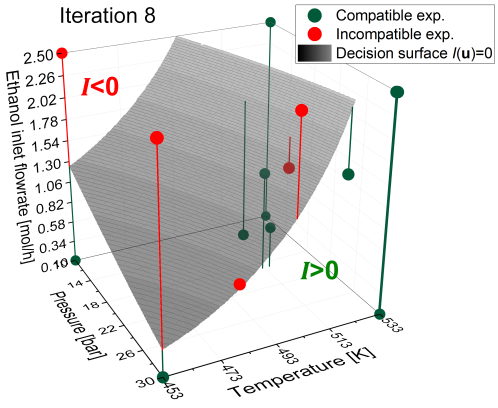
One shall notice from (2.29) that the binary variables $\beta_i\ \forall\ i = 1,...,N$ are functions of the parameter values θ. The value computed by the solution of the MBDM problem in (2.27) may change significantly when additional samples are introduced in the objective function. As a consequence, the classification of a specific sample may change in the course of the experimental campaign. The MBDM estimator selects for the fitting the subset of samples Y' which maximises the objective function $\mathscr{L}_{DM}$. As one can see from Table 4.2, the number of fitted samples can either increase or decrease in the course of the experimental campaign. This is explained by the tendency of MBDM to give fitting priority to samples



(a)



(b)



(c)

Figure 4.3: Considered experimental design space defined by pressure, temperature and ethanol inlet flowrate at different iterations of the model identification procedure implementing a constrained MBDoE: (a) after the collection of the 8 preliminary samples; (b) after the collection of 4 designed samples; (c) after the collection of 8 designed samples. Green dots and red dots represent observed compatible (i.e. $\hat{\beta}_j = +1$) and incompatible (i.e. $\hat{\beta}_j = -1$) experimental conditions respectively, according to the labelling computed by MBDM. The grey surface at $I(\varphi) = 0$ represents the optimal boundary for the domain of model reliability computed by the Support Vector Classifier.

with small associated residuals. A qualitative sketch is proposed in Figure 4.4 to illustrate this scenario. In Figure 4.4a two generic samples in the dataset, i.e. sample i and j are fitted by MBDM because their squared residuals are below the maximum threshold $N_y c^2$. In this condition, both samples give a positive contribution to the objective function $\mathscr{L}_{DM}$. In Figure 4.4b, an additional sample k is collected and included in the dataset. When MBDM is applied, sample i and sample j are excluded from the fitting to give fitting priority to sample k. In fact, sample k alone brings a higher contribution to the objective function $\mathscr{L}_{DM}$ than samples i and j together. Hence, it may happen that a highly compatible sample is included in the dataset and a number of previously compatible samples are excluded from the objective function to give fitting priority to the new sample.

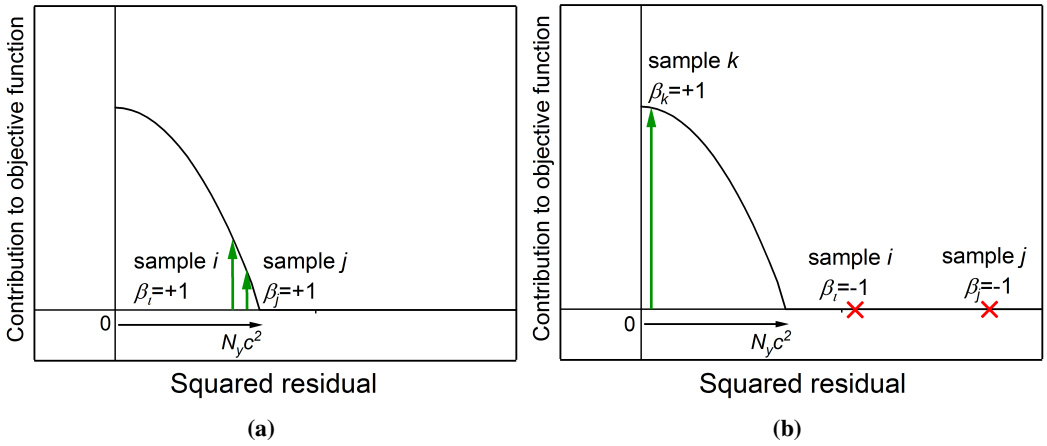


Figure 4.4: A possible effect of the MBDM approach when it is applied online. In the plots, green arrows indicate a positive contribution of a sample to the objective function $\mathscr{L}_{DM}$. The fitting of multiple samples with large residuals (a) may contribute less than a single sample with small residual (b) to the objective function. The fitting of sample k in (b) results in a better optimum than the fitting of samples i and j together. When sample k is included in the dataset, sample i and j are excluded from the fitting by MBDM.

The identification of the model in the MBDM case required the collection of 16 samples, while only 11 samples were used for parameter fitting. This is due to the inaccurate approximation of the model reliability domain computed by SVM when the number of performed experiments in the training set is limited. This inaccuracy can lead to the design of samples at conditions where the model is inaccurate even if a constrained MBDoE approach is employed. These model-incompatible samples are subsequently discarded by MBDM and ignored in the model identification process. However, the accuracy of the SVM classification improves in the course of the experimental design campaign and does not prevent the ultimate identification of an approximated model that is accurate within its

reliability domain. The model instance identified in the MBDM case is characterised by a better fitting compared to the model identified in the ML case. The final sum of squared residuals in the MBDM case is $\chi^2_Y = 45.57$, which is within the acceptable range assumed in the goodness-of-fit test. In different words, the model identified in the MBDM case is not falsified by the data used for the estimation of its parameters.

The precise estimation of the parameters in the MBDM case requires the fitting of 11 samples, while only 8 samples were required to identify the model in the ML case. In the ML case, the information from all the available samples is used for the estimation. In the MBDM case, only the information from samples collected within the domain of reliability is exploited. Regions of the design space within the model reliability domain may be associated with suboptimal levels of Fisher information. If the estimation of parameters in the MBDM case must be performed fitting less informative data, a higher number of samples must be fitted to achieve the same level of precision as in the ML case.

4.3.2 Case study 2: methanol oxidation on silver

4.3.2.1 Experimental setup and data set

A microreactor platform is available to perform kinetic experiments for the identification of a kinetic model of methanol oxidation on silver. A schematic diagram of the device is given in Figure 4.5. The reactor chip was constructed from a silicon wafer through photolithography and deep reactive ion etching. A thin layer of silver was sputtered on the bottom of the microchannel obtaining a catalyst film 78.1 mm in length. Mass flow controllers were used to inject the gaseous mixture consisting of methanol, oxygen, water and helium (added as inert carrier). A detailed description of the experimental setup is available in the literature (Cao and Gavriilidis, 2005). The independent conditions that can be manipulated in the system are: the temperature T [K] of the microreactor; the flowrate F [ml min^{-1}] of the gaseous mixture at the inlet; molar fractions of methanol, oxygen and water in the inlet mixture, i.e., $y^{IN}_{CH_3OH}$, $y^{IN}_{O_2}$ and $y^{IN}_{H_2O}$ respectively.

A dataset Y consisting of $N = 13$ samples of the outlet composition was obtained performing 13 steady-state experiments varying one factor at time. Each sample in Y includes $N_y = 6$ measurements, namely the outlet molar fraction of methanol, oxygen, water, formaldehyde, hydrogen and carbon dioxide. A summary of the experimental conditions investigated for the collection of the dataset is given in Table 4.3. The complete dataset is reported in Appendix E.

Table 4.3: Experimental conditions investigated in the available dataset. The volumetric flowrate F is referred to standard conditions. Helium, used as inert carrier, represents the remaining molar fraction at the inlet.

Sample number	T [K]	F* [ml min^{-1}]	$y^{IN}_{CH_3OH}$	$y^{IN}_{O_2}$	$y^{IN}_{H_2O}$
1-3	783	29.1-73.1	0.0996	0.0414	0.0754
4-7	733-826	50.9	0.1468	0.0975	0.2293
8-10	765-826	93.9	0.1469	0.0980	0.2296
11-13	800-900	54.5	0.2590	0.1064	0.2122

* at temperature $T = 273.15$ K; pressure $P = 101325$ Pa.

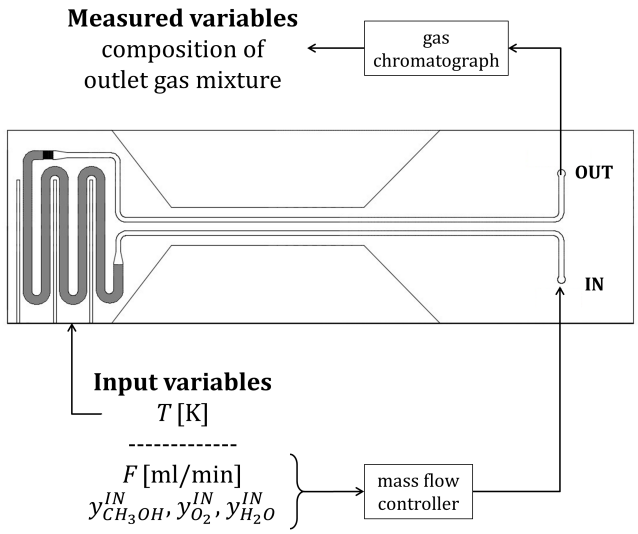


Figure 4.5: Schematic representation of the microreactor chip and setup. The grey-coloured area in the microchannel represents the sputtered silver catalyst film.

4.3.2.2 Approximated model

The section of the microchannel occupied by the silver catalyst film is modelled as an ideal plug-flow reactor. Isothermal conditions are assumed along the channel (i.e. the energy balance is omitted), and diffusion phenomena are completely neglected. The generic form of the mass balance is given in (4.9), where N_C and N_R represent the number of components and the number of reactions respectively, C_i is the species concentration expressed in mol m^{-3}, z represents the axial coordinate of the channel in m, v is the flow velocity along z expressed in m s^{-1}, ν_{ij} is the stoichiometric coefficient of the i-th component in the j-th reaction and r_j is the rate associated with the j-th reaction, expressed in mol m^{-3} s^{-1}.

$$v\frac{dC_i}{dz} = \sum_{j=1}^{N_R} \nu_{ij} r_j \quad \forall \quad i = 1,...,N_C \tag{4.9}$$

Andreasen et al. (2003) formulated a micro-kinetic model for the reaction based on surface science studies. From this micro-kinetic model, the same authors identified the presence of two limiting steps in the oxidation. The first step is the intermediate methoxy decomposition that results in the generation of formaldehyde and hydrogen. The second step is the intermediate formate decomposition that results in the generation of carbon dioxide and hydrogen. From these considerations, a simplified kinetic model which involves only two lumped reactions was derived (Andreasen et al., 2005). As in other works available in the literature, the two-reaction model proposed by Andreasen *et al.* is employed adding a third reaction of hydrogen oxidation to account for the low amounts of hydrogen detected at the outlet of the reactor (Galvanin et al., 2015). The stoichiometry and kinetics of the assumed reactions are

$$\begin{aligned} &\text{Reaction 1:} \quad \mathrm{CH_3OH} + \tfrac{1}{4}\mathrm{O_2} \leftharpoons \mathrm{CH_2O} + \tfrac{1}{2}\mathrm{H_2} + \tfrac{1}{2}\mathrm{H_2O} \\ &\text{Reaction 2:} \quad \mathrm{CH_2O} + \tfrac{1}{2}\mathrm{O_2} \leftharpoons \mathrm{H_2} + \mathrm{CO_2} \\ &\text{Reaction 3:} \quad \mathrm{H_2} + \tfrac{1}{2}\mathrm{O_2} \rightarrow \mathrm{H_2O} \end{aligned} \tag{4.10}$$

A total of $N_C = 6$ species are considered in the approximated kinetics, i.e., methanol, oxygen, water, formaldehyde, hydrogen and carbon dioxide. The rates of the three reactions are given in (4.11), where R is the ideal gas constant, A_j and E_{aj} (with $j = 1, ..., 3$) represent pre-exponential factors and activation energies of the Arrhenius type rate constants.

$$\begin{aligned} r_1 &= A_1 e^{-\frac{E_{a1}}{RT}} \frac{C_{\mathrm{CH_3OH}} C_{\mathrm{O_2}}^{0.25}}{C_{\mathrm{H_2O}}^{0.5}} \\ r_2 &= A_2 e^{-\frac{E_{a2}}{RT}} \frac{C_{\mathrm{CH_2O}} C_{\mathrm{O_2}}^{0.5}}{C_{\mathrm{H_2}}^{0.5}} \\ r_3 &= A_3 e^{-\frac{E_{a3}}{RT}} C_{\mathrm{H_2}} C_{\mathrm{O_2}}^{0.5} \end{aligned} \tag{4.11}$$

An instance for the kinetic parameters was available from previous kinetic investigations, conducted on a different setup (Quaglio et al., 2019). The values are reported in Table 4.4. The silver catalyst considered in this work went through a different fabrication history compared to the catalyst used in previous works. Hence, one shall not expect the parameter instance given in Table 4.4 to be representative for the catalyst employed in this case study. The different kinetic behaviour between different silver catalyst types is assumed to derive from a different density of active sites on the film surface. Following this assumption,

only the pre-exponential factors of the catalytic reactions shall be tuned on the available data set Y. The catalyst promotes the partial oxidation of methanol and the oxidation of formaldehyde, i.e., reaction 1 and reaction 2. Evidence reported in the literature suggests that reaction 3 occurs slowly on the catalyst surface (Schubert et al., 1994; Dokuchits et al., 2012). Therefore, in this case study, the catalytic effect of silver on reaction 3 is neglected, i.e., it is assumed that a different density of active sites on the catalyst surface does not influence the kinetic rate of hydrogen oxidation. Thus, the kinetic constants A_3, E_{a1}, E_{a2} and E_{a3} are fixed to the values given in Table 4.4 and only A_1 and A_2 are treated as the parameters requiring re-estimation, i.e., $\theta = [A_1, A_2]$.

Table 4.4: Instance for the kinetic parameters obtained from previous kinetic studies.

Parameter	Unit	Value
A_1	$[(\mathrm{mol\,m^{-3}})^{0.25}\mathrm{s^{-1}}]$	$5.33 \cdot 10^{11}$
A_2	$[\mathrm{s^{-1}}]$	$1.03 \cdot 10^{7}$
A_3	$[(\mathrm{mol\,m^{-3}})^{-0.5}\mathrm{s^{-1}}]$	$1.07 \cdot 10^{4}$
E_{a1}	$[\mathrm{J\,mol^{-1}}]$	$1.42 \cdot 10^{5}$
E_{a2}	$[\mathrm{J\,mol^{-1}}]$	$9.02 \cdot 10^{4}$
E_{a3}	$[\mathrm{J\,mol^{-1}}]$	$1.83 \cdot 10^{4}$

4.3.2.3 Objective and methods

Since the model presented in Section 4.3.2.2 was derived by a number of simplifying hypotheses, its identification requires both the quantification of the unknown parameters $\theta = [A_1, A_2]$ through the fitting of dataset Y, and the identification of the reliability domain associated with the estimated parameters. As in the previous case study (see Section 4.3.1), two scenarios are presented:

ML case The model is identified using a standard Maximum Likelihood estimator (Bard, 1974).

MBDM case The model is identified employing the methodology presented in Section 4.2. The following settings are adopted for the application of the proposed approach:

- *MBDM settings.* The MBDM problem is formulated as in (2.27). The measured molar fractions in a sample are assumed to be affected by Gaussian noise with covariance $\Sigma_y = 3 \cdot 10^{-3} \cdot \mathbf{I}$. The MBDM tolerance is set at $c = 3.0$. This is equivalent to treating as outliers measurements with an associated residual above the 3 standard deviations range.

- *SVM settings.* A SVM is employed to identify a reliability map $I(\varphi)$ in the experimental design space. Two cases are considered. In *Case 1*, the model is assumed to be weak at describing certain ranges of temperature and inlet fraction of methanol while it is assumed to be reliable on the other experimental conditions. The SVM machine is therefore trained assuming a bi-dimensional input space $\Phi = (T, y^{IN}_{CH_3OH})$. In *Case 2*, the model is considered weak in representing the system in broad ranges of temperature and inlet fraction of water, but reliable on other experimental conditions. The SVM machine is then trained on the reduced input space $\Phi = (T, y^{IN}_{H_2O})$.

The experimental design step is not considered in this case study. The ML and MBDM problems are solved employing respectively the solvers MAXLKHD and CVP_SS in gPROMS ModelBuilder 4.1 (PSE gPROMS, 2017). In the MBDM case, the model reliability map is identified using the tool for support vector classification implemented in the Python package *scikit-learn* (Pedregosa et al., 2011). In the present case study, the hyperparameters of the SVM are set *a priori*. The Gaussian kernel in (4.1) is employed with $\gamma = 0.2$; being the experimental conditions in the training set normalised, this corresponds to having a characteristic decay length equal to 20% of the explorable range in any direction of Φ. The regularisation constants are set to the default value implemented in *scikit-learn*.

4.3.2.4 Results and discussion

The model parameters were fitted to the dataset using both a conventional ML estimator and MBDM. The parameter estimates are reported in Table 4.5 with the associated t-value statistics and the sum of squared residuals χ^2_Y (for additional details on the performed statistical tests see Section 2.4.4). As one can see, all the computed parameters are statistically satisfactory, but the estimates obtained in the two cases are significantly different. The reason is that in the MBDM case, some of the binary variables β were switched to -1 to satisfy the MBDM conditions in (2.29), excluding some samples from the parameter estimation problem. In Table 4.6, the binary variables $\hat{\beta}$ computed by MBDM are given for all the experiments together with the conditions investigated for the collection of each sample. The samples 4, 8, 12 and 13 (i.e., the samples with $\hat{\beta} = -1$) were labelled by MBDM as incompatible with the modelling assumptions. The parity plot in Figure 4.6a shows the distribution of the residuals achieved by the candidate model if the ML method is employed (i.e., if the whole dataset is fitted). In Figure 4.6b, the residuals associated with the fitted

data in the MBDM case (i.e., only the residuals associated with samples 1-3, 5-7 and 9-11) are reported. The distributions of the normalised residuals associated with the ML method and with the MBDM method are plotted in Figure 4.7a and Figure 4.7b respectively. From a comparison of the plots in Figure 4.6 and the bar charts in Figure 4.7 one can see that the application of MBDM led to the identification of a model with improved fitting capabilities. The exclusion of experiments 4, 8, 12 and 13 results in a significant reduction of the χ^2_Y from 1247.2 in the ML case to 180.3 in the MBDM case.

Table 4.5: Parameter estimates and related statistics: t-value and sum of squared residuals χ^2; with conventional ML estimator and MBDM estimator.

Method	Estimates $\hat{\theta} = [A_1, A_2]$	t-values*	t_{ref}	χ^2_Y
ML	$[5.66 \cdot 10^{12}, 7.33 \cdot 10^{7}]$	$[19.51, 15.39]$	1.66	1247.2
MBDM	$[3.98 \cdot 10^{12}, 6.16 \cdot 10^{7}]$	$[14.63, 11.26]$	1.67	180.3

*a t-value higher than t_{ref} indicates satisfactory parameter precision.

Table 4.6: Experimental conditions investigated in the catalytic microreactor and binary variables $\hat{\beta}$ computed by MBDM. Samples with $\hat{\beta} = -1$ were not considered for the estimation of the kinetic parameters.

Sample	T [K]	F* [ml min^{-1}]	$y^{IN}_{CH_3OH}$	$y^{IN}_{O_2}$	$y^{IN}_{H_2O}$	$\hat{\beta}$
1	783	73.1	0.0996	0.0414	0.0754	+1
2	783	41.7	0.0996	0.0414	0.0754	+1
3	783	29.1	0.0996	0.0414	0.0754	+1
4	733	50.9	0.1468	0.0975	0.2293	-1
5	765	50.9	0.1468	0.0975	0.2293	+1
6	796	50.9	0.1468	0.0975	0.2293	+1
7	826	50.9	0.1468	0.0975	0.2293	+1
8	765	93.9	0.1469	0.0980	0.2296	-1
9	796	93.9	0.1469	0.0980	0.2296	+1
10	826	93.9	0.1469	0.0980	0.2296	+1
11	800	54.5	0.2590	0.1064	0.2122	+1
12	850	54.5	0.2590	0.1064	0.2122	-1
13	900	54.5	0.2590	0.1064	0.2122	-1

* at temperature $T = 273.15$ K; pressure $P = 101325$ Pa.

The classified samples are used to train a SVM classifier and compute a model reliability map $I(\varphi)$ in the form of (4.2). The map obtained for *Case 1* is represented in Figure 4.8a in the input subspace defined by temperature and inlet fraction of methanol. Regions of the input space at $I(\varphi) > 0$ (bright regions in the plot) identify conditions at which the model is expected to provide a good representation of the reacting system. Conversely, conditions at

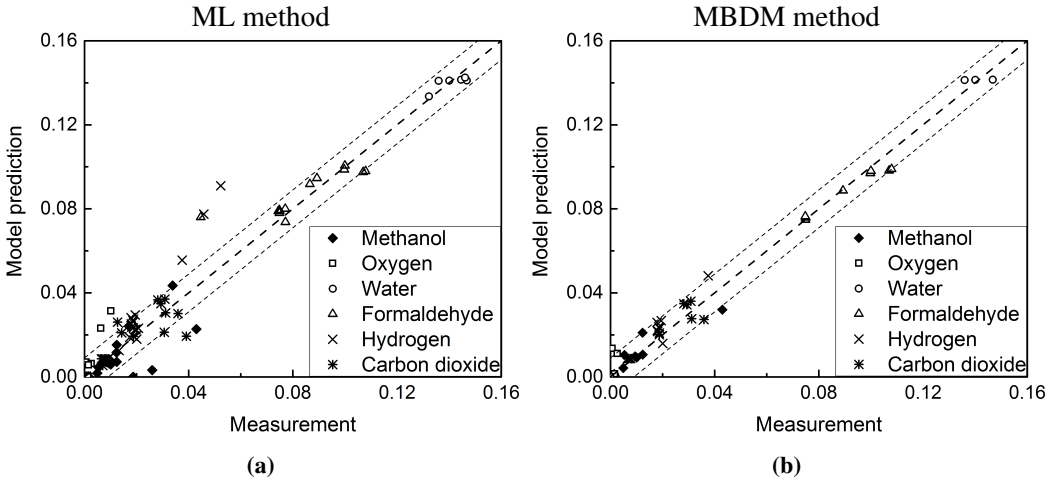


Figure 4.6: Parity plot comparing measurements against model predictions: (a) if a conventional ML estimator is employed; (b) if MBDM is adopted. In (b) only experimental data with $\hat{\beta} = +1$ are reported.

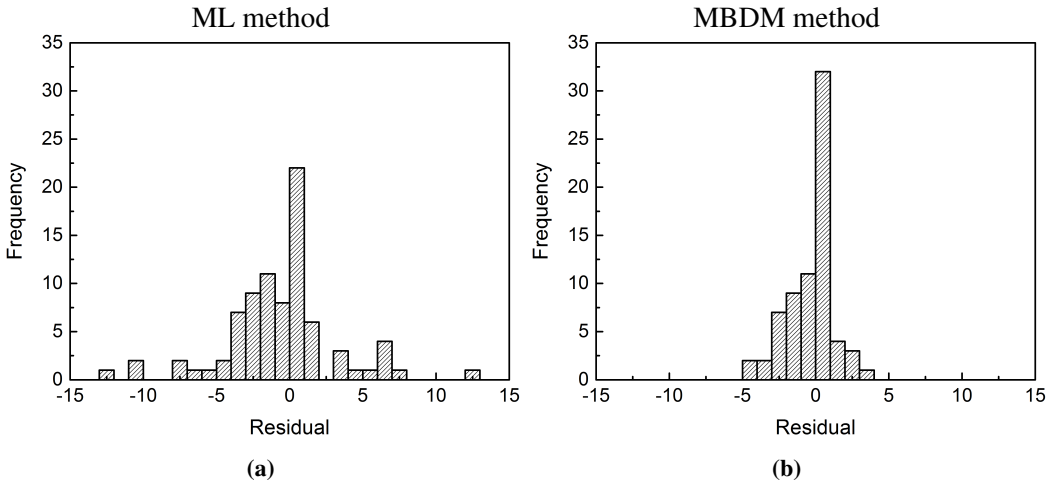


Figure 4.7: Distribution of the normalised residuals: (a) if a conventional ML estimator is employed; (b) if MBDM is adopted. In (b) only residuals associated with the experimental data with $\hat{\beta} = +1$ are reported.

$I(\varphi) < 0$ (dark regions in the plot) are considered too close to samples that were previously classified as incompatible with the candidate model. The reliability map identified in *Case 2* is plotted in Figure 4.8b in the input subspace defined by temperature and fraction of water at the inlet. In this case study, the reliability map was computed in two cases where the experimental design space is two-dimensional. However, maps of reliability may be easily computed selecting different sets of training variables, possibly including more than two experimental conditions.

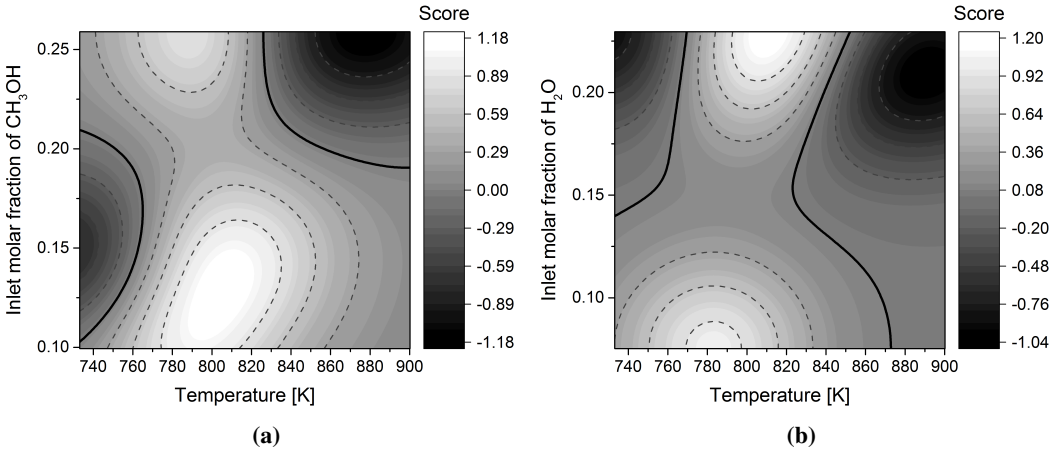


Figure 4.8: Score of decision functions identified training the SVM with two different sets of experimental conditions: (a) temperature and methanol fraction at the inlet; (b) temperature and water fraction at the inlet. Solid black lines represent contours at $I(\varphi) = 0$.

4.3.3 Computational times and problem size

In the methanol oxidation case study, the identification of the model required the estimation of $N_\theta = 2$ kinetic parameters fitting $N = 13$ samples. The MBDM problem was formulated as an MINLP with 2 continuous variables (i.e., the kinetic parameters) and 13 binary variables (i.e., the binary switchers β_i with $i = 1, ..., N$). The problem was solved using the OAERAP solver (Adjiman et al., 1998) implemented in gPROMS and required 4.29 s.

In the ethanol dehydrogenation case study, the approximated model involved $N_\theta = 4$ kinetic parameters. Its identification was performed online in a simulated experimental campaign where the MBDM problem was solved with a number of samples ranging from $N = 8$ to $N = 16$. Both MBDM and constrained experimental design problems were solved using the SLSQP routine implemented in the package *SciPy* (Jones et al., 2001). The computational times associated with the solution of the MBDM and constrained experimental design problems are reported in Table 4.7. The CPU times associated with the solution of the MBDM problem ranged from 29.32 s to 238.60 s. It is recognised that the CPU time associated with the application of MBDM depends on the initial guess and the number of optimisation variables. Furthermore, it is recognised that the employment of more advanced optimisation routines for MINLP problems, e.g. the solver OAERAP (Adjiman et al., 1998) or BARON (Sahinidis, 1996), may significantly improve the convergence rate of the optimisation at the MBDM stage. The implementation in Python of a package for solving MBDM problems using robust MINLP solvers will be the objective of future work.

In some cases, the design of a single experiment was performed, i.e., $N_{sp} = 1$. If the

predicted information from a single experiment were not sufficient to achieve the desired parameter precision, a design involving multiple experiments up to $N_{sp} = 3$ was performed (see Section 4.3.1.3). The number of optimisation variables in the constrained experimental design problems is $3 \cdot N_{sp}$. In fact, 3 experimental conditions (ethanol inlet flowrate, pressure and temperature) are optimised independently for each designed experiment. The design problems are solved using the SLSQP routine implemented in *SciPy* (Jones et al., 2001). One can observe from Table 4.7 that the design of a single experiment with 3 optimisation variables required a CPU time between 21.73 s and 28.09 s; the design of $N_{sp} = 2$ experiments required between 65.95 s and 174.23 s. Only after the collection of sample 12, a design with $N_{sp} = 3$ experiments was performed and required 489.32 s.

If the identification of the model is performed online, the time required to design the following experiments should be substantially shorter than the sampling frequency allowed in the experimental setup. The CPU time associated with the experimental design stage may be reduced by designing a small number N_{sp} of experiments simultaneously and/or by employing more efficient optimisation solvers.

Table 4.7: Ethanol dehydrogenation on copper. Computational times [s] associated with the solution of the MBDM problem and the design of constrained experiments from the collection of sample 8 to the collection of the last sample 16. N_{sp} indicates the number of experiments that were simultaneously designed. N/A indicates that no experimental design was performed because a lower number of designed experiments was found adequate to meet the desired statistical quality of the parameter estimates.

	Algorithm runtime [s]			
Collected samples	MBDM problem	Constrained exp. design $N_{sp} = 1$	$N_{sp} = 2$	$N_{sp} = 3$
8	66.78	25.95	163.68	N/A
9	92.12	24.86	97.59	N/A
10	103.29	25.45	N/A	N/A
11	94.70	27.43	65.95	N/A
12	194.93	22.75	174.23	489.32
13	126.48	12.98	N/A	N/A
14	67.15	28.09	N/A	N/A
15	238.60	21.73	N/A	N/A
16	29.32	N/A	N/A	N/A

4.4 Final remarks

The identification of an approximated model is a problem of multi-objective nature that requires *i*) the estimation of the model parameters by fitting experimental data and *ii*) the

computation of the domain of reliability associated with the estimated parameter values. A framework for the identification of approximated models was presented in this Chapter.

If the model structure is approximated, one shall not expect the model to be accurate across the entire range of observable experimental conditions. Data collected at conditions where the model is inaccurate shall be treated as outliers and neglected for the purpose of parameter estimation. In the proposed framework, model parameters are estimated using a Model-Based Data Mining (MBDM) method, which is derived from robust regression theory. MBDM produces two outputs: *1*) it classifies the available samples in terms of low or high model residuals compared with a user-defined accuracy tolerance c; *2*) it returns the maximum likelihood estimate associated with the fitting of only the samples with low residuals (i.e., the samples that are in agreement with the model predictions).

The classified samples returned by MBDM are used to train a Support Vector Machine (SVM) classifier with Gaussian kernel. The training of SVM results in the computation of a decision function I, which quantifies the expected model accuracy across the space of experimental conditions. If one is willing to enhance the precision on the model parameters by fitting additional data, the research of optimal conditions through MBDoE methods shall be bounded to regions of the design space space where $I > 0$, i.e., where the model is expected to provide a good fitting.

Notice that the inclusion of additional samples in the dataset does not necessarily lead to the computation of different parameter estimates (the new data may in fact be classified as outliers by MBDM). However, the inclusion of new samples in the dataset always results in an update of the model reliability map I. Since the SVM classifier in (4.2) is influenced by all the available samples, its score will increase or decrease in the neighbourhood of the conditions associated with the new sample depending on the labelling computed by MBDM.

The mapping of the design space provided by SVM is influenced by the choice of the kernel function as well as the values of the associated hyperparameters (see Section 4.2.2). Furthermore, an accurate SVM classification requires the availability of a relatively abundant and distributed training set. Especially at the beginning of the experimental activity, the number of samples may be limited and the classification may be poor. However, notice that an inaccurate classification would only impact the efficiency of the method (i.e., the number of samples required to identify the model) and not the eventual outcome. An initially inaccurate reliability map may lead to the collection of incompatible samples in re-

gions of the design space that are classified as reliable. However, the accuracy of the SVM classifier increases as the experimental activity proceeds and does not prevent the ultimate identification of an approximated model that is accurate within its reliability domain.

The domain of reliability computed by SVM tends to approximate the range of conditions in which residuals are within the range of c standard deviations of measurement noise. In extreme cases, if the tolerance c is too small, the reliability domain may be extremely narrow and the information available in the constrained design space may not be sufficient for estimating precisely the parameters, i.e., the model may not be identifiable within its reliability domain. In such cases, one may choose to relax the MBDM tolerance c and expand the model reliability domain. Otherwise, if high model accuracy is a fundamental requirement, one may prefer to test alternative model structures.

When the approximated model is identified, one may use the reliability map to check if the model is appropriate for a specific application. As an example, the scientist may want to employ the model to identify an optimal range of conditions to maximise the performance of a process under study. This range of conditions will be denoted as the domain of application of the model. If the domain of application lies within the domain of reliability, as in Figure 4.9a, the model is expected to be reliable at the conditions of interest. In such case, the approximated model may be adequate to optimise the process under study. Conversely, if the domain of application is not contained within the domain of reliability, as in Figure 4.9b, the model shall not be trusted. In this scenario, one may employ an experimental design approach where additional samples are designed in the domain of application with the aim of improving the model performance in that region of the design space. This possible modelling path will be explored in future research activities. Alternatively, one may conclude that the available model is inadequate for optimising the process and the model structure should be modified embracing the available experimental evidence. In the following Chapter, a systematic approach is proposed to diagnose model misspecification and inform the scientist on how an approximated model structure can be improved.

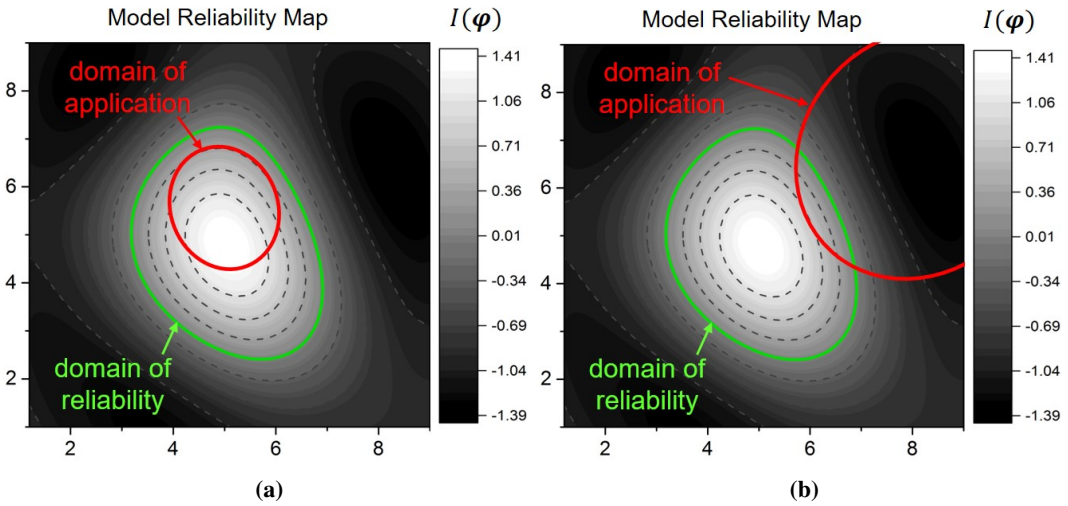


Figure 4.9: Graphical comparison between the domain of model reliability, i.e., the range of conditions where the model is expected to be accurate, and the domain of application for the model, i.e., the range of conditions where the scientist wants the model to be accurate. (a) the domain of application is within the domain of reliability, i.e., the model is appropriate for the specific application. (b) the domain of application is not contained within the domain of reliability, i.e., the model is not reliable in the conditions of interest and a different model structure should be preferred.

Chapter 5

Diagnosis of model misspecification

Part of this Chapter is adapted from the following articles:

Quaglio M., Fraga E. S. and Galvanin F., Statistical diagnosis of process-model mismatch by means of the Lagrange Multipliers test, *Proceedings of the 29th European Symposium on Computer Aided Process Engineering*, 2019, pp. 679-684

Quaglio M., Fraga E. S., Galvanin F., A diagnostic procedure for improving the structure of approximated kinetic models, *Computers & Chemical Engineering*, 2019 (in press)

The author of this Thesis contributed to the above articles by developing the main novel ideas, implementing the simulations, and writing a significant part of the text. Hence, the author retains the right to include the articles in this Thesis since it is not published commercially and the journals are referenced as the original source.

5.1 Introduction

Whenever a model is falsified by data, its mathematical structure should be modified embracing the available experimental evidence. A framework based on maximum likelihood inference is illustrated in this work for diagnosing model misspecification and improving the structure of approximated models. In the proposed framework, statistical evidence provides a measure to justify a modification of the model structure, namely the removal of irrelevant parameters and/or the evolution of relevant parameters into state-dependent expressions. A tailored Lagrange multipliers test (see Section 2.7) is proposed to detect which model parameters are expected to improve model fitting quality the most should they be evolved into state-dependent quantities.

5.2 Proposed methodology

A dataset $Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ consisting of N samples of $\mathbf{y}$ is assumed to be available for modelling the kinetic behaviour of a process under study. As in previous Chapters, the symbol φ_i denotes the set of experimental conditions adopted for the collection of the i-th sample in Y. A framework for kinetic model building is introduced in Figure 5.1. The framework begins with the construction of an approximated kinetic model. Recall the generic model structure

$$\begin{aligned} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t, \theta) &= \mathbf{0} \\ \hat{\mathbf{y}} &= \mathbf{h}(\mathbf{x}, \mathbf{u}, t, \theta) \end{aligned} \tag{2.1}$$

The identification of the model involves the estimation of a set of N_θ parameters θ. It is assumed that the model satisfies the requirement for practical identifiability given the available dataset Y (see Section 2.5). Hence, the model parameters θ can be uniquely estimated by fitting the dataset Y.

The framework then involves the following sequential steps:

1. *Parameter estimation.* The model parameters are fitted to the available dataset using a maximum likelihood approach (Bard, 1974).

2. *Goodness-of-fit test.* The adequacy of the model in representing the dataset is assessed with a two-tailed test on the goodness-of-fit (Silvey, 1975). A two-tailed test is employed to detect modeling errors either when model residuals are too small or too large compared with the level of measurement noise present in the system. The test has three possible outcomes:

   (a) *Passed.* The model is not falsified and it is adequate for representing the dataset. There is no evidence for justifying a change in the model structure.

   (b) *Failed for over-fitting.* The model may involve parameters that are unnecessary for representing the process. If over-fitting is detected, one shall proceed by performing a Wald test (Wald, 1943) for parameter significance. Unnecessary parameters are removed from the model structure and the procedure is repeated from step 1.

   (c) *Failed for under-fitting.* The model structure does not capture the underlying dynamics of the physical system. A tailored Lagrange multipliers test (Silvey, 1959) is proposed in this work as a tool for measuring the statistical evidence to

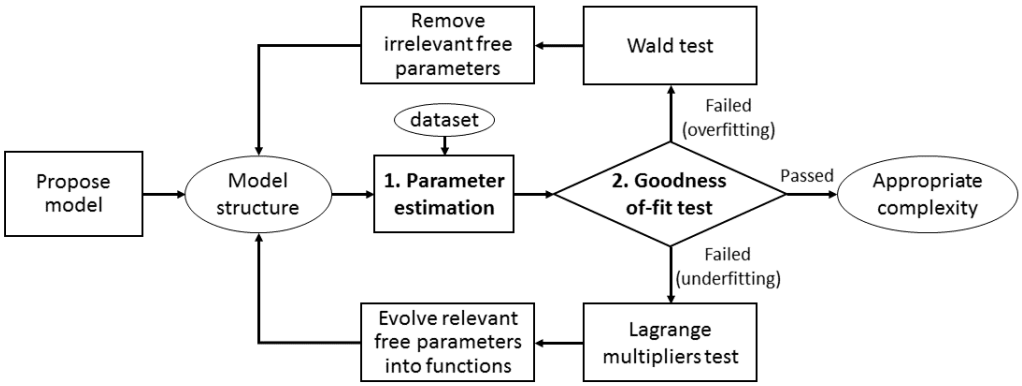


Figure 5.1: Proposed framework for kinetic model building. In the proposed approach, statistical tests are performed to diagnose model misspecification and to support the scientist in the improvement of misspecified model structures.

disprove the hypothesis that a given parameter is a state-independent quantity. The model structure is evolved by substituting the parameter (or parameters) with highest associated evidence with an opportune function of the states and the procedure is repeated from step 1.

The illustrated procedure is further detailed in the following subsections assuming that the model is falsified for under-fitting. Particular emphasis is given to the description of the Lagrange multipliers test, which is proposed to diagnose model descriptive limits and inform on which parameters should be considered for evolution into state-dependent expressions. The model evolution step in the procedure will be discussed in Chapter 6.

5.2.1 Parameter estimation

Model parameters θ are estimated with a maximum likelihood approach (see Section 2.4.3). Recall, the log-likelihood function $\mathscr{L}$ is

$$\begin{aligned}\mathscr{L}(Y|\theta) &= -\frac{N}{2}[N_y \ln(2\pi) + \ln(\det(\Sigma_y))] \\ &- \frac{1}{2}\sum_{i=1}^{N}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta_1,...,\theta_{N_\theta})]^T \Sigma_y^{-1}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta_1,...,\theta_{N_\theta})]\end{aligned} \tag{2.7}$$

The maximum likelihood estimate $\hat{\theta} = [\hat{\theta}_1,...,\hat{\theta}_{N_\theta}]^T$ is computed by maximizing the unconstrained log-likelihood function

$$\hat{\theta} = \arg\max_{\theta} \mathscr{L}(Y|\theta) \tag{2.8}$$

The maximum likelihood estimate satisfies the unconstrained likelihood equations

$$\nabla \mathscr{L}(Y|\hat{\theta}) = \mathbf{0} \tag{2.9}$$

5.2.2 Goodness-of-fit test

Once the model parameters are fitted to the available dataset, the adequacy of the model is checked with a goodness-of-fit test (see Section 2.4.4.1) based on a two-tailed χ^2 test. Under the hypothesis of the proposed model being *exact*, the sum of normalized squared residuals χ^2_Y is distributed as a χ^2 distribution with degree of freedom $N \cdot N_y - N_\theta$

$$\chi^2_Y = \sum_{i=1}^{N} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\hat{\theta})]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\hat{\theta})] \sim \chi^2_{N \cdot N_y - N_\theta} \tag{2.12}$$

In this work, a two-tailed χ^2 test with significance $\alpha = 90\%$ is used. A significance of 90% in a two-tailed test represents a typical value assumed in statistical inference (Devore, 2010). If the statistic χ^2_Y lies between the 5% and the 95% percentiles of the χ^2 distribution, the model is considered as an adequate representation of the physical system. Whenever χ^2_Y is below the 5% percentile, the model is falsified for over-fitting. If χ^2_Y is above the 95% percentile, the model is falsified for under-fitting.

5.2.3 Lagrange multipliers test

When the model is under-fitting, a significant discrepancy between experimental observations and model predictions is observed. It is assumed that a reduction of the discrepancy (and eventually its elimination to the limit of measurement noise) can be achieved by evolving a certain model parameter into an opportune function of the state variables. A statistical test is proposed to diagnose model misspecification by challenging the hypothesis that a given parameter θ_i is a state-independent constant. The proposed test aims at diagnosing whether it is appropriate to assume a specific model component as a free parameter or whether a significant improvement in the model fitting quality is expected should that parameter be replaced with a function of the state variables. Without loss of generality, the test is detailed assuming that the parameter under diagnosis is the first parameter, i.e., $\theta_i = \theta_1$. The competing hypotheses under test are:

Null hypothesis H_0. θ_1 and $\theta_j \ \forall \ j \neq 1$ are all state-independent constants.

Alternative hypothesis H_a. θ_1 is a state-dependent function and $\theta_j \ \forall \ j \neq 1$ are state-

independent constants.

The parameter estimation problem is formulated under the assumptions that θ_1 is a function g of the experimental conditions, i.e., $\theta_1 = g(\varphi)$, and $\theta_j \ \forall \ j \neq 1$ are fixed coefficients. One shall notice that no assumption on the functional form of g is required to perform the test. The $N \times 1$ parameter array Θ_d (subscript d stands for diagnosis) is defined as $\Theta_d = [\theta_{1,1}, ..., \theta_{1,N}]^T$ where the i-th element in the array represents the value of g at the experimental conditions φ_i, i.e., $\theta_{1,i} = g(\varphi_i) \ \forall \, i = 1, ..., N$. The log-likelihood function $\mathscr{L}_d$ is constructed under parametrization Θ_d

$$\begin{aligned} \mathscr{L}_d(Y|\Theta_d) &= -\frac{N}{2}[N_y \ln(2\pi) + \ln(\det(\Sigma_y))] \\ &- \frac{1}{2}\sum_{i=1}^{N}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta_{1,i}, \hat{\theta}_2, ..., \hat{\theta}_{N_\theta})]^T \Sigma_y^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta_{1,i}, \hat{\theta}_2, ..., \hat{\theta}_{N_\theta})] \end{aligned} \tag{5.1}$$

In (5.1), the i-th element in the sum is a function of parameter $\theta_{1,i}$ only. The other model parameters are set equal to their maximum likelihood value and treated as fixed constants in the test, i.e., $\theta_j = \hat{\theta}_j \ \forall \ j \neq 1$. In words, it is assumed that the parameter under diagnosis does not interact with the other model parameters.

The set of $N-1$ functions $\mathbf{s}$ is defined as

$$\mathbf{s} = [\theta_{1,1} - \theta_{1,2}, ..., \theta_{1,N-1} - \theta_{1,N}]^T \tag{5.2}$$

The null and alternative hypotheses are then formalised mathematically as the presence/absence of an $N-1$ set of constraints for the functions $\mathbf{s}$ as follows

$$\begin{aligned} H_0 : \quad & \mathbf{s} = \mathbf{0} \\ H_a : \quad & \mathbf{s} \neq \mathbf{0} \end{aligned} \tag{5.3}$$

Notice that the imposition of constraints $\mathbf{s} = \mathbf{0}$ is equivalent to assuming that g is a constant function that is independent from the experimental conditions φ. The constrained maximum likelihood estimate $\hat{\Theta}_d = [\hat{\theta}_{1,1}, ..., \hat{\theta}_{1,N}]^T$ is obtained by maximizing the log-likelihood function $\mathscr{L}_d$ under constraints $\mathbf{s} = \mathbf{0}$.

$$\begin{aligned} \hat{\Theta}_d = \arg\max_{\Theta_d} & \mathscr{L}_d(Y|\Theta_d) \\ \text{s.t.} \quad & \mathbf{s} = \mathbf{0} \end{aligned} \tag{5.4}$$

Under constraints $\mathbf{s} = \mathbf{0}$ all the elements in $\hat{\Theta}_d$ are equal to the unconstrained maximum likelihood estimate for parameter θ_1, i.e., $\hat{\theta}_{1,i} = \hat{\theta}_1 \; \forall i = 1, ..., N$. The constrained maximum likelihood estimate $\hat{\Theta}_d$ also satisfies the set of constrained maximum likelihood equations

$$\begin{aligned} \nabla \mathscr{L}_d(Y|\hat{\Theta}_d) + \nabla \mathbf{s} \hat{\boldsymbol{\alpha}} &= \mathbf{0} \\ \mathbf{s} &= \mathbf{0} \end{aligned} \tag{5.5}$$

where $\hat{\alpha}$ is the $N-1 \times 1$ array of Lagrange multipliers associated to the constraints. As demonstrated by Aitchison and Silvey (1958) and Silvey (1959), under the null hypothesis being true, the Lagrange multipliers statistic $\xi_d(\theta_1)$ is asymptotically distributed as a χ^2 distribution with degree of freedom equal to the number of constraints (i.e., $N-1$) as shown in the following equation

$$\xi_d(\theta_1) = \hat{\boldsymbol{\alpha}}^T \nabla \mathbf{s}^T \mathbf{H}_d^{-1} \nabla \mathbf{s} \hat{\boldsymbol{\alpha}} \sim \chi^2_{N-1} \tag{5.6}$$

In (5.6), $\mathbf{H}_d$ represents the $N \times N$ expected Fisher information matrix for the model under parametrization Θ_d, which is well approximated by the following expression under null hypothesis conditions

$$\mathbf{H}_d = \sum_{i=1}^{N} \nabla \hat{\mathbf{y}}_i(\hat{\theta}_{1,i}) \Sigma_y^{-1} \nabla \hat{\mathbf{y}}_i(\hat{\theta}_{1,i})^T \tag{5.7}$$

Notice that the solution of the constrained maximum likelihood equations (5.5) is not required to compute the statistic ξ_d. In fact, ξ_d may be directly computed as a function of the log-likelihood gradient evaluated setting $\Theta_d = \hat{\Theta}_d$ as follows

$$\xi_d = \nabla \mathscr{L}_d(Y|\hat{\Theta}_d)^T \mathbf{H}_d^{-1} \nabla \mathscr{L}_d(Y|\hat{\Theta}_d) \sim \chi^2_{N-1} \tag{5.8}$$

and does not require the evaluation of the Lagrange multipliers $\hat{\alpha}$ (Rao, 1948). In this work, the Lagrange multipliers statistic is computed according to the expression in (5.8).

The illustrated approach for constructing the statistic $\xi_d(\theta_1)$, associated with parameter θ_1, is repeated for all model parameters obtaining the set of statistics $\xi_d(\theta_i) \; \forall i = 1, ..., N_\theta$. A measure of model misspecification, namely a Model Modification Index (MMI), is defined

as

$$\mathrm{MMI}(\theta_i) = \frac{\xi_d(\theta_i)}{\chi^2_{N-1}(95\%)} \quad \forall\, i = 1,...,N_\theta \tag{5.9}$$

The MMI represents a ratio between a Lagrange multipliers statistic and the 95% percentile of the χ^2 distribution with degree of freedom $N-1$. A MMI larger than 1 indicates that the null hypothesis is falsified by a χ^2 test with 95% of significance. If $\mathrm{MMI}(\theta_i) > 1$, one shall expect a significant improvement in the model fitting quality if parameter θ_i were evolved into a state-dependent function. Conversely, if $\mathrm{MMI}(\theta_i) < 1$ there is no statistical evidence for justifying the evolution of parameter θ_i into a function of the states. The MMI quantifies the expected rate of increase in the log-likelihood function associated with an infinitesimal relaxation of the constraint $\mathbf{s} = \mathbf{0}$. Hence, if the null hypothesis is falsified for more than one parameter, one shall expect a more significant improvement in the model fitting quality if the parameters with the highest MMI were evolved. A MMI-based diagnosis of model misspecification may be performed by using a radar chart as in Figure 5.2. In the example, the model involves $N_\theta = 5$ parameters. The MMIs associated with parameters θ_1, θ_2 and θ_4 are below 1, i.e., there is no evidence to justify the evolution of these parameters into functions. $\mathrm{MMI}(\theta_3)$ and $\mathrm{MMI}(\theta_5)$ are above 1. Hence, the analysis suggests that a significant improvement of the fitting may be achieved by evolving either θ_3 or θ_5 into some function. Nonetheless, since $\mathrm{MMI}(\theta_3) > \mathrm{MMI}(\theta_5) > 1$, a more significant improvement in the fitting is expected from the evolution of θ_3, compared with θ_5.

The MMI formulated in (5.9) does not consider the possible interaction between the parameter under diagnosis and the other model parameters. It is recognised that if parameter interaction is considered, the computation of the MMI may not be possible unless an appropriate experimental design is adopted for the collection of the dataset Y. This is due to the possible singularity of the information matrix, which must be invertible to compute the Lagrange multipliers statistic. In this work, the study of experimental design criteria for MMI-based model misspecification diagnosis will not be considered. Nonetheless, a multivariate MMI that considers the effect of parameter interaction is formulated in Appendix H where necessary conditions for the application of a multivariate MMI-based diagnosis are also derived.

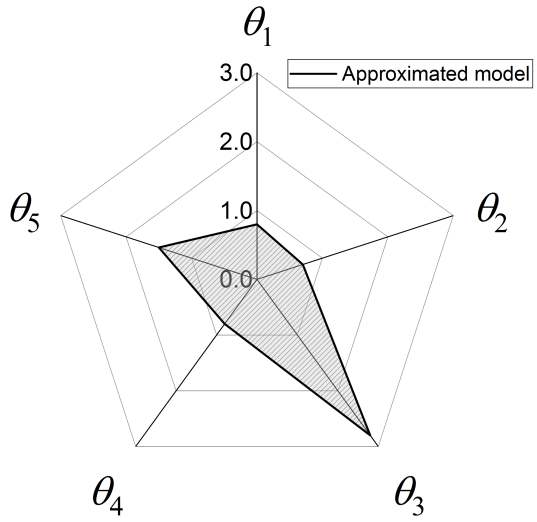


Figure 5.2: A possible approach for visualising the MMIs is through a radar chart. In this example, the model involves 5 parameters under diagnosis. The MMI associated with parameter θ_3 and θ_5 are above 1. Hence, the analysis suggests that a significant improvement in the model fitting quality may be achieved if either θ_3 or θ_5 were evolved into some opportune state-dependent function.

5.3 Case studies

In this section, two simulated case studies are presented to demonstrate the Lagrange multipliers test proposed in Section 5.2.3 for the diagnosis of model misspecification. In case study 1 (Section 5.3.1), the test is applied to the diagnosis of under-fitting in an approximated model of baker's yeast growth (Asprey and Macchietto, 2000). In case study 2 (Section 5.3.2), the MMI-based approach is employed to diagnose under-fitting in an approximated model of glucose-insulin interaction (Bergman et al., 1981). In case study 2, the sensitivity of the MMIs to a change in the experimental design and in the system noise is assessed. The numerical results presented in this section were obtained with Python 3.5 (Python Core Team, 2018).

5.3.1 Case study 1: baker's yeast growth model

5.3.1.1 System model

The considered system is a cultivation of yeast in a fed-batch bioreactor. The system kinetics are assumed to be described by the following set of differential and algebraic equations:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = (r - u_1 - \theta_4)x_1 \tag{5.10}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \tag{5.11}$$

$$r = \frac{\theta_1 x_2}{\theta_2 x_1 + x_2} \tag{5.12}$$

where $x_1(t)$ [g L^{-1}] is the yeast concentration and $x_2(t)$ [g L^{-1}] is the substrate concentration. The kinetic behaviour of the system is expressed as a function of two system inputs, namely the dilution factor u_1 [h^{-1}] and the substrate concentration in the feed u_2 [g L^{-1}]. In the system model, the yeast growth rate r obeys a Contois-type kinetic law. The system model involves a set of $N_\theta = 4$ parameters θ whose values are $\theta^* = [0.310, 0.180, 0.550, 0.050]^T$.

5.3.1.2 Approximated model

It is assumed that the scientist does not know the functional form of the system model and proposes an approximated model structure. The approximated model includes equations (5.10) and (5.11) with a Monod-type kinetic law

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = (r - u_1 - \theta_4)x_1 \tag{5.10}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \tag{5.11}$$

$$r = \frac{\theta_1 x_2}{\theta_2 + x_2} \tag{5.13}$$

The approximated model and the system model differ in the functional form of the rate expression. The element $\theta_2 x_1$ appearing at the denominator in (5.12) is modeled as a state independent parameter, i.e. θ_2, in the denominator of the approximated rate law (5.13). The identification of the approximated model requires the estimation of a set of $N_\theta = 4$ parameters θ.

5.3.1.3 Objective and Methods

The objective in this case study is to diagnose model misspecification in the approximated model of yeast growth presented in Section 5.3.1.2 using an approach based on the computation of the MMIs. It is assumed that an array $\mathbf{y} = [x_1, x_2]^T$ of system states can be sampled in the simulated experiments. Samples of $\mathbf{y}$ are assumed to be corrupted by uncorrelated Gaussian measurement noise with covariance $\Sigma_y = 2.5 \cdot 10^{-3}\mathbf{I}$.

A full factorial experimental design with four dynamic experiments is assumed with two levels for the dilution factor, i.e. $u_1 = \{0.05, 0.20\}$ h^{-1}, and two levels for the substrate concentration in the feed, i.e. $u_2 = \{5.0, 35.0\}$ g L^{-1}. In each experiment, 7 samples of $\mathbf{y}$ are collected at sampling times $t_s = \{3.0, 6.0, 9.0, 12.0, 15.0, 18.0, 21.0\}$ h. The initial conditions for the differential variables are the same in all the experiments, i.e., $x_1(0) = 1.0$ g L^{-1} and $x_2(0) = 0.01$ g L^{-1}. A dataset Y is generated in-silico by integrating the system model presented in Section 5.3.1.1 and adding random Gaussian noise with covariance Σ_y. The complete dataset is reported in Appendix F.

The dataset is fitted both with the system model and with the approximated model. The MMIs are then computed for both model structures. The procedure is performed on both models to assess the behaviour of the MMIs both in the presence of appropriate and inappropriate modelling assumptions.

5.3.1.4 Results and discussion

When the system model is used to fit the dataset, the sum of squared residuals is $\chi^2_Y = 61.32$, which lies within the acceptable range assumed for the two-tailed goodness-of-fit test with 90% of significance, i.e. $36.44 < \chi^2_Y < 69.83$. The test suggests that there is no evidence for evolving the model structure. The MMIs associated with the system model are reported in Table 5.1 and plotted in the radar chart in Figure 5.3a for visualisation purposes. All the

Table 5.1: Baker's yeast system. Goodness-of-fit test and model modification index for all model parameters. Results are presented both for the system model and for the approximated model structure.

Model structure	Goodness-of-fit test				MMIs associated to
	$\chi^2(5\%)$	χ^2_Y	$\chi^2(95\%)$	Outcome	$[\theta_1, \theta_2, \theta_3, \theta_4]$
System model	36.44	61.32	69.83	Passed	[0.67, 0.74, 0.77, 0.77]
Approximated model	36.44	2210.37	69.83	Failed for under-fitting	[16.98, 47.08, 11.90, 18.58]

MMIs associated with the parameters in the system model are below 1. Hence, there is no evidence to justify the evolution of any parameter in the system model.

The parameter set involved in the approximated model is estimated by fitting the dataset. As one can see from Table 5.1, the approximated model is falsified by the goodness-of-fit test. More specifically, a sum of squared residuals $\chi^2_Y = 2210.37$ larger than the χ^2 value at 95% of significance highlights the presence of significant under-fitting. The approximated model should be modified by replacing some parameter with an opportune function of the state-variables. The MMIs associated with the model parameters are plotted in Figure 5.3b. The MMI is larger than 1 for all model parameters. Hence, a significant improvement of the fitting quality is expected if any of the model parameters were evolved. The highest MMI is associated to θ_2, i.e. $\text{MMI}(\theta_2) = 47.08$. The scientist may then focus on choosing an opportune state-dependent function to replace parameter θ_2 in the approximated model. In practice, the *exact* model structure is unknown, nonetheless, in this simulated case study one can appreciate that it is possible to make the approximated model indistinguishable from the system model by replacing parameter θ_2 with the functional form $\theta_2 x_1$. Hence, the MMI-based analysis correctly highlights that a major improvement on the model fitting may be achieved by evolving θ_2.

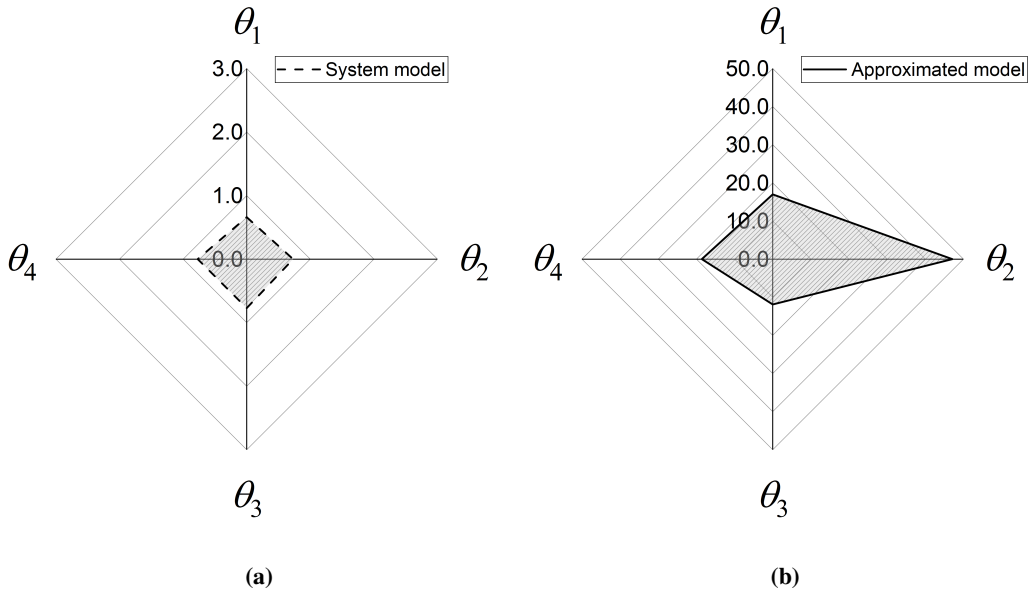


Figure 5.3: Baker's yeast system. Model Modification Index associated with the model parameters of (a) the system model and (b) the approximated model.

5.3.2 Case study 2: glucose-insulin interaction model

5.3.2.1 System model

The physical system of interest in this case study is the glucose-insulin regulatory system of a healthy test subject with basal glucose concentration $G_b = 93.0$ mg dL^{-1}. The system dynamics are described by the following set of equations (Bergman et al., 1981)

$$\frac{\mathrm{d}G}{\mathrm{d}t} = -\theta_1(G - G_b) - \theta_2 X G \tag{5.14}$$

$$\frac{\mathrm{d}X}{\mathrm{d}t} = -\theta_3 X + I \tag{5.15}$$

$$\mathrm{IDR} = \max[0, \quad \theta_4(G - \theta_5)t] \tag{5.16}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \mathrm{IDR} - \theta_6 I \tag{5.17}$$

where $G(t)$ [mg dL^{-1}] is the plasma glucose concentration, $X(t)$ [μU min mL^{-1}] represents the insulin action term associated with the remote insulin receptor (Bergman et al., 1981, 1979; Zeleznik and Roth, 1978; Insel et al., 1975), $I(t)$ [μU mL^{-1}] is the plasma insulin concentration and IDR represents the insulin delivery rate as a function of glucose concentration in plasma (Toffolo et al., 1980). The system model involves a set θ of $N_\theta = 6$ parameters. The values of the system parameters associated with the test subject are $\theta^* = [2.96 \cdot 10^{-2}, 6.51 \cdot 10^{-6}, 1.86 \cdot 10^{-2}, 5.36 \cdot 10^{-3}, 9.09 \cdot 10^{1}, 2.3 \cdot 10^{-1}]^T$.

5.3.2.2 Approximated model

The physiologist proposes an approximated model for the system which involves the following set of differential and algebraic equations (Bergman et al., 1981)

$$\frac{\mathrm{d}G}{\mathrm{d}t} = -\theta_1(G - G_b) - \theta_2 X \tag{5.18}$$

$$\frac{\mathrm{d}X}{\mathrm{d}t} = -\theta_3 X + I \tag{5.15}$$

$$\mathrm{IDR} = \max[0, \quad \theta_4(G - \theta_5)t] \tag{5.16}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \mathrm{IDR} - \theta_6 I \tag{5.17}$$

The system model and the approximated model differ in the functional form of equations (5.14) and (5.18), which describe the glucose concentration in plasma. The nonlinear term $-\theta_2 X G$ appearing in the system equation (5.14) is modelled as a linear term, i.e.

$-\theta_2 X$, in the approximated model equation (5.18). The approximated model structure involves a set of $N_\theta = 6$ kinetic parameters θ.

5.3.2.3 Objective and Methods

It is assumed that G, I and X may be sampled from the patient during an intravenous glucose tolerance test (IVGTT). The protocol assumed for the IVGTT is the same adopted by Bergman et al. (1981), where 23 samples are collected from the test subject in the course of a 182.0 min assay.

The model identification approach proposed in Section 5.2 is applied to diagnose model misspecification in the approximated model structure presented in Section 5.3.2.2. Four different cases are considered to assess the sensitivity of the MMIs to a change in the experimental design (i.e., different sets of measured state variables and different initial conditions of the test subject) and to a change in the level of measurement noise in the system. In all the illustrated cases, the dataset is generated in-silico by integrating the system model described in Section 5.3.2.1 and adding random measurement noise to the measured states. The simulated datasets analysed in this case study are reported in Appendix G. The cases are summarized in Table 5.2 and further described in the following list.

Case A. A single IVGTT is performed at initial conditions $G(0) = 298.0$ mg dL^{-1}, $I(0) = 333.0$ μU mL^{-1}, $X(0) = 0.0$ μU min mL^{-1}. The sample includes measurements for G and I, i.e. $\mathbf{y} = [G, I]^T$. A low level of uncorrelated, Gaussian system noise is assumed with standard deviations 1.0 mg dL^{-1} for measurements of G and 1.5 μU mL^{-1} for measurements of I. The dataset is reported in Table G.1 (only measured values for G and I are considered in this case).

Case B. Same as Case A, but measuring also the insulin action X, i.e. $\mathbf{y} = [G, I, X]^T$. Measurement noise for X is characterized by a standard deviation of 10.0 μU min mL^{-1}. The dataset is reported in Table G.1.

Table 5.2: Glucose-Insulin interaction system. Summary of cases considered in the study.

Case ID	IVGTT number	Measured variables	Measurement noise
A	1	G, I	low
B	1	G, I, X	low
C	2	G, I	low
D	1	G, I	high

Case C. Same as Case A, but with an additional IVGTT performed at initial conditions $G(0) = 276.0$ mg dL^{-1}, $I(0) = 69.0$ μU mL^{-1}, $X(0) = 0.0$ μU min mL^{-1}. Data associated with the additional IVGTT are reported in Table G.2.

Case D. Same as Case A, but assuming high system noise with standard deviations 5.0 mg dL^{-1} for G and 7.5 μU mL^{-1} for I. The dataset considered in this case is reported in Table G.3.

For all the cases, the parameters of both system model and approximated model are fitted to the data. The goodness-of-fit test is then performed and the MMIs are computed for both the system and the approximated model structure. As in case study 1, the MMIs are evaluated for both models to assess their behaviour in the presence of both an appropriate and an inappropriate set of modelling hypotheses.

Table 5.3: Glucose-insulin interaction system. Goodness-of-fit test and model modification index for all model parameters. Results are presented both for the system model and for the approximated model structure in the different considered cases.

Case	Model	Goodness-of-fit test				MMIs associated to
ID	structure	$\chi^2(5\%)$	χ^2_Y	$\chi^2(95\%)$	Outcome	$[\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]$
A	System model	26.51	42.05	55.57	Passed	[0.61, 0.61, 0.61, 0.70, 0.66, 0.80]
	Approximated model	26.51	97.8	55.57	Failed for under-fitting	[2.20, 2.19, 2.20, 1.18, 1.17, 1.40]
B	System model	45.57	64.34	82.57	Passed	[0.69, 0.59, 0.64, 0.80, 0.79, 0.71]
	Approximated model	45.74	128.75	82.57	Failed for under-fitting	[2.17, 2.15, 0.68, 0.85, 0.86, 0.76]
C	System model	65.62	85.99	108.64	Passed	[0.81, 0.80, 0.79, 0.69, 0.69, 0.74]
	Approximated model	65.62	334.71	108.64	Failed for under-fitting	[4.23, 4.27, 4.27, 2.39, 1.91, 2.63]
D	System model	26.51	36.96	55.57	Passed	[0.45, 0.45, 0.46, 0.72, 0.65, 0.64]
	Approximated model	26.51	49.17	55.57	Passed	[0.81, 0.83, 0.84, 0.77, 0.73, 0.91]

5.3.2.4 Results and discussion

Numerical results for the goodness-of-fit test and computed MMIs are reported in Table 5.3. As one can see from Table 5.3, in all cases, the system model passes the goodness-of-fit test and its associated MMIs are always below 1, suggesting that there is no evidence to justify an evolution of the model structure. The approximated model is falsified for under-fitting in Cases A-C. In Case D, the approximated model is not falsified due to an excessive level of

system noise. The MMIs associated with the system model are plotted in the radar charts in Figure 5.4 (dotted lines) together with the MMIs associated with the approximated model (solid lines) for a visual comparison.

The MMIs associated with the approximated model in Case A are plotted in Figure 5.4a (solid line). As one can see from Figure 5.4a, all the MMIs associated with the approximated model are higher than 1. The MMIs associated with θ_1, θ_2 and θ_3 are the largest with a value around 2.20. The analysis suggests that the most significant improvement in the model fitting quality may be achieved by evolving any of these parameters. In practice there is uncertainty on how these parameters could be evolved. Nonetheless, since the system model is known in this case study, it is possible show that an opportune evolution of parameters θ_1, θ_2 or θ_3 in the approximated model can make it indistinguishable from the system model structure. The discrepancy between approximated and system model structures vanishes if parameter θ_1 were evolved into $\theta_1 + \theta_2 X(G-1)/(G-G_b)$ or θ_2 were evolved into $\theta_2 G$. It is also possible to make the approximated model indistinguishable from the system model by evolving θ_3. In fact, by evolving θ_3, it is possible to modify the behaviour of variable X in order to compensate for the absence of state G in the addend $-\theta_2 X$ in (5.18).

A change in equation (5.15) has the potential of improving the fitting quality for variables G and I without causing a degradation in the fitting quality for X. In fact, variable X is not observed in Case A. The observed under-fitting vanishes if parameter θ_3 evolves into the function

$$\theta_1 + \theta_3 - \frac{GI}{X} + \frac{I}{X} - \frac{\theta_1 G_b}{G} + \frac{\theta_2 X}{G} \tag{5.19}$$

In Case B, measurements of X are included in the log-likelihood function. The MMIs associated with the approximated model in Case B are plotted in Figure 5.4b (solid line). In Case B, only the MMIs of parameters θ_1 and θ_2 are above 1. The Lagrange multipliers test does not suggest the evolution of parameters $\theta_3 - \theta_6$. Parameters $\theta_3 - \theta_6$ are involved in the correctly specified equations (5.15) and (5.17), and their evolution is not expected to improve the fitting quality.

In Case C, the inclusion in the log-likelihood function of an additional IVGTT causes an increase of all MMIs with respect to Case A. The MMIs associated with the approximated kinetic model in Case C are plotted in Figure 5.4c (solid line). As in Case A, also in Case C the state X is not observed and the Lagrange multipliers test suggests that a major

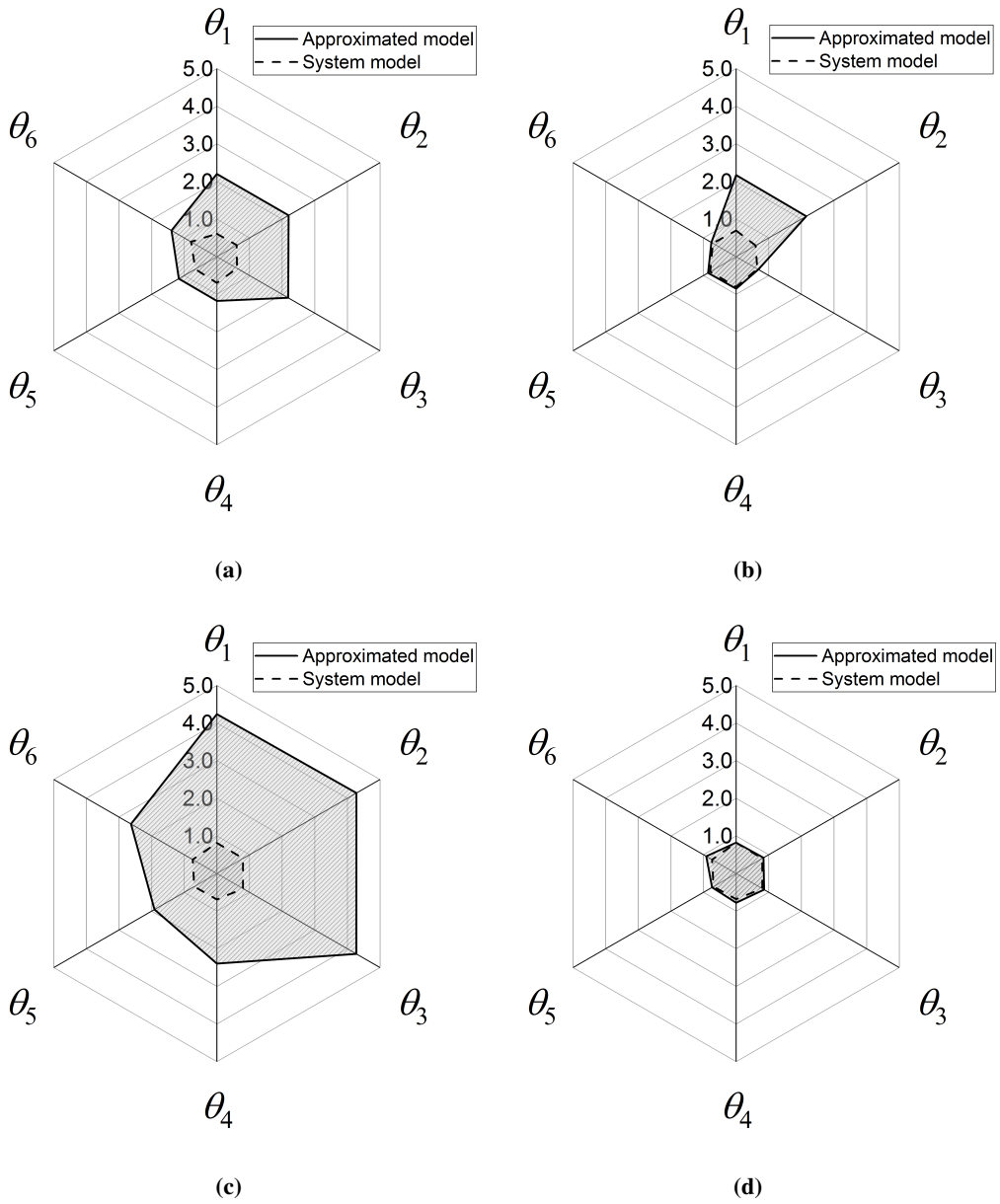


Figure 5.4: Glucose-Insulin interaction system. Model Modification Indexes for all model parameters. (a) Case A: One performed IVGTT; G, I observed variables; low system noise. (b) Case B: One performed IVGTT; G, I, X observed variables; low system noise. (c) Case C: Two performed IVGTTs; G, I observed variables; low system noise. (d) Case D: One performed IVGTT; G, I observed variables; high system noise. For all Cases, MMIs are plotted for the approximated model (solid line) and for the system model (dotted line).

model improvement may be achieved by evolving parameter θ_1, θ_2 or θ_3. A less significant improvement may be achieved by evolving parameters θ_4, θ_5 or θ_6. As in Case A, a change in the correctly specified equations (5.16) and (5.17) may benefit the fitting quality for variable G, but it would cause a degradation in the fitting quality of I.

In Case D, the approximated model does not fail the goodness-of-fit test. The high system noise in Case D prevents the falsification of the modelling hypothesis and there is no evidence to modify the model structure. The MMIs in Case D, plotted in Figure 5.4d (solid line) are all below 1, suggesting that no parameter should be evolved.

5.3.3 Computational times and problem size

Quantitative information on the size of the model identification problems considered in the case studies is reported in Table 5.4 together with a min-max algorithm runtime range associated with the computation of the MMIs. In the baker's yeast case study, the approximated model consisted of 2 ODEs involving $N_\theta = 4$ parameters. The dataset Y consisted of $N = 28$ samples where each sample involved $N_y = 2$ measured quantities, namely the biomass concentration x_1 and the substrate concentration x_2. The computation of the MMIs associated with the approximated model required only few seconds of CPU time and never more than 26.0 s.

In the case study on the glucose-insulin regulatory system, the approximated model structure involved 3 ODEs and included $N_\theta = 6$ parameters. Furthermore, as one can see in Section 5.3.1.2, a discontinuity associated with the IDR dynamics was present in the approximated model structure. The dataset Y in Case A consisted of $N = 23$ samples, each involving $N_y = 2$ measured quantities, namely the glucose concentration in plasma G and the insulin concentration in plasma I. The computation of the MMIs required around 25.0 s per parameter as shown in Table 5.4.

The computational time associated with the evaluation of the MMIs is not significantly different in the two case studies despite the difference in the number of model equations and model parameters. In fact, it is recognised that the computational time is primarily influenced by the number of samples N in the dataset Y. The computation of the MMIs requires the evaluation of the $N \times 1$ gradient of the log-likelihood function $\mathscr{L}_d(Y|\hat{\boldsymbol{\theta}}_d)$, and the computation and inversion of the $N \times N$ Fisher information matrix $\mathbf{H}_d$. However, it is also observed that the amount of samples available in kinetic modelling studies is typically small. It is therefore expected that in most practical cases it will be possible to compute the

MMIs with a limited employment of computational resources.

Table 5.4: Problem size and algorithm runtime [s] for case study 1, i.e., the Baker's yeast system, and case study 2, i.e., the glucose-insulin regulatory system. The number of ODEs, parameters N_θ and samples N considered in the respective case studies are reported. The runtime is given in the form of a min-max range.

Case	Problem size			Runtime for
study	ODEs	Parameters N_θ	Samples N	MMI evaluation [s]
1	2	4	28	22.61 - 25.56
2 (Case A)	3	6	23	24.90 - 25.01

5.4 Final remarks

A diagnostic procedure based on maximum likelihood inference is illustrated in this Chapter to support scientist in the improvement of approximated kinetic model structures. In the proposed model building framework, modifications in the model structure are justified and supported by experimental evidence. When the model is over-fitting, model parameters that are *irrelevant* for representing the data are removed from the model structure. A Wald test is used to determine which model parameters one shall omit from the model. When the model is in conditions of under-fitting, *relevant* model parameters are evolved into more state-dependent functions. A tailored Lagrange multipliers test is proposed in this work to determine which model parameters one shall consider to substitute with state-dependent expressions.

The proposed Lagrange multipliers test does not require the definition of alternative model structures or superstructures. In fact, the test aims at disproving the null hypothesis that a given model parameter under diagnosis is a state-independent constant. A model modification index (MMI) is introduced as a function of a Lagrange multipliers statistic. Parameters with the highest MMI are those that are expected to improve the model fitting quality the most if they were replaced with state-dependent functions. When the MMI is below unity there is scarce evidence for justifying an alteration of the parameter. The test was demonstrated in a number of simulated cases with a baker's yeast growth model and with a model of glucose-insulin interaction. It is shown that, in the presence of moderate system noise, the MMIs correctly highlight the parameters that are primarily associated with model misspecification. When the system noise increases, the falsification of an incorrect modelling hypothesis for under-fitting becomes increasingly challenging and a decrease in

the MMIs is observed. When the system noise is excessive, the falsification of an incorrectly specified model structure with a finite dataset may be impractical and the computed MMIs decrease below unity, suggesting that there is no evidence to justify the evolution.

The MMI represents a scalar measure of model misspecification that accounts for system measurement noise, model residuals and parameter sensitivities. Nonetheless, the MMI formulated in this Chapter neglects the interaction between the parameter under diagnosis and the other free model parameters. A multivariate MMI that considers parameter interaction is formulated in Appendix H, where it is shown that a multivariate MMI-based analysis is possible only if an appropriate experimental design is adopted. In fact, in the presence of an inappropriate experimental design, the Fisher information matrix may be non-invertible and it may not be possible to compute the Lagrange multipliers statistic considering parameter interaction. Future work shall focus on the study of sufficient conditions for an experimental design to advocate a multivariate MMI-based diagnosis of model misspecification.

In the next Chapter, additional tests will be developed to support the scientist in the selection of appropriate functional forms to replace critical model parameters in underfitting models.

Chapter 6

Evolution of kinetic model structures

Part of this Chapter is adapted from the following articles:

Quaglio M., Fraga E. S., Galvanin F., A diagnostic procedure for improving the structure of approximated kinetic models, *Computers & Chemical Engineering*, 2019 (in press)

Quaglio M., Fraga E. S., Galvanin F., The evolution of approximated kinetic model structures, *Proceedings of the 2019 AIChE Annual Meeting*, 2019

The author of this Thesis contributed to the above articles by developing the main novel ideas, implementing the simulations, and writing a significant part of the text. Hence, the author retains the right to include the articles in this Thesis since it is not published commercially and the journals are referenced as the original source.

6.1 Introduction

A statistical tool for diagnosing model misspecification in under-fitting models was proposed in Chapter 5. A Model Modification Index (MMI) was defined as a function of a Lagrange multipliers statistic to detect which model parameters are likely to hide state dependencies. Whenever a high MMI is computed for a given model parameter, a significant improvement in the model fitting quality is expected should that parameter be replaced with an opportune function of the state variables. In this Chapter, an Effect Relevance Index (ERI) is introduced as a computationally cheap heuristic to quantify the relevance of a candidate effect for the evolution of model parameters into state-dependent functions. An analysis on the ERIs may inform the scientist on which effects are the most important to consider in the definition of state-dependent expressions to evolve model parameters and improve the model fitting performance.

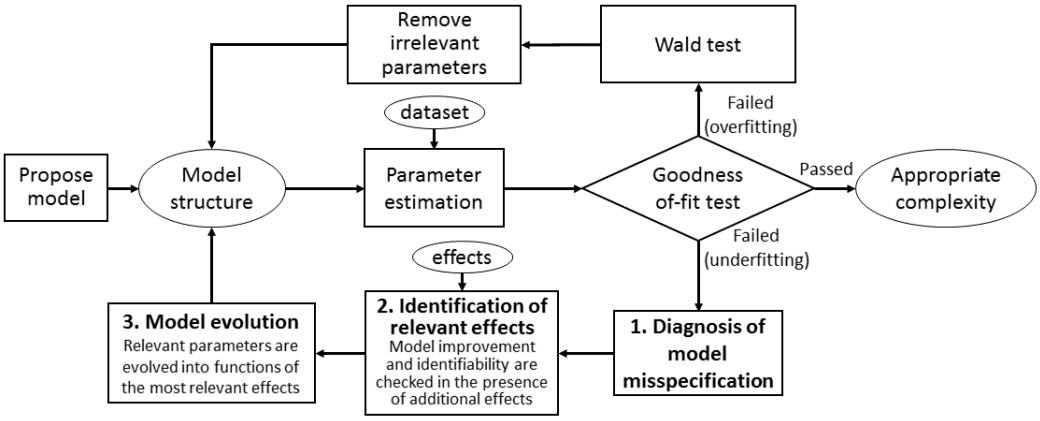


Figure 6.1: Proposed framework for kinetic model building. In the proposed approach, statistical tests are performed to diagnose model misspecification and to support the scientist in the improvement of misspecified model structures. Particular emphasis in the framework is given to the improvement of a model structure when under-fitting is detected.

6.2 Proposed methodology

A setup is available for studying the dynamics of a physical system of interest. It is assumed that a dataset Y in the form (2.6) is available to identify a kinetic model of the system. The model building approach illustrated in Section 5.2 is employed for the construction and identification of a kinetic model. The framework is re-proposed in Figure 6.1, where the boldface blocks represent the main focus in the present Chapter. The scientist proposes an approximated model in the usual form (2.1). This model represents the initial model structure and will be denoted as M_0.

$$M_0 : \begin{cases} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t, \theta) = \mathbf{0} \\ \hat{\mathbf{y}} = \mathbf{h}(\mathbf{x}, \mathbf{u}, t, \theta) \end{cases} \tag{2.1}$$

The maximum likelihood estimate for the model parameters $\hat{\theta} = [\hat{\theta}_1, ..., \hat{\theta}_{N_\theta}]^T$ is obtained by fitting the dataset Y using a maximum likelihood approach. It is assumed that the model is falsified for under-fitting by the two-tailed goodness-of-fit test. The procedure then involves the following sequential steps:

1. *Diagnosis of model misspecification.* At this stage, the scientist performs a diagnosis of model misspecification based on the computation of the Model Modification Indexes (MMIs) for all the model parameters (see Section 5.2.3). Parameters with the highest MMIs are those that are expected to improve the model fitting quality the most should they be evolved into state-dependent functions. The scientist selects the

parameters with the highest MMI for evolution into state-dependent functions.

2. *Identification of relevant effects.* The scientist may not know a-priori which functional forms should be chosen to evolve the parameters with the highest MMI. The scientist shall proceed by proposing a set of candidate effects that may be relevant for the construction of such functional forms. An Effect Relevance Index (ERI) is then computed from a Lagrange multipliers statistic (see Section 2.7) to quantify the relevance of each effect on the candidate parameters selected for evolution. The computation of the ERI requires the inversion of the expected Fisher information matrix in an extended parameter space. Hence, the ERI for a given effect is evaluated only if practical model identifiability requirements are respected in the presence of the extended parametrisation (for more information on practical identifiability see Section 2.5).

3. *Model evolution.* The model structure is evolved by replacing the parameters selected for evolution with opportune functions of the effects with the highest ERI.

The aforementioned steps represent an iteration in the proposed model building procedure for improving the structure of under-fitting models. These main stages will be further detailed in the following subsections. Particular emphasis will be given to the approach proposed to compute the ERIs. The use of the ERIs will also be demonstrated in simulated case studies highlighting the strengths of the approach and discussing its limitations.

6.2.1 Diagnosis of model misspecification

Model misspecification is diagnosed by computing the Model Modification Index (MMI) for all the model parameters in the set θ. A detailed discussion on the computation and application of the MMIs for model diagnosis is given in Chapter 5. If the MMI is above 1 for some model parameter, a significant improvement in the model fitting quality is expected, should that parameter be evolved into a state-dependent function. If the MMI is higher than 1 for multiple parameters, one shall expect a more significant improvement in the model fitting quality if the parameters with the highest MMI were evolved. A MMI-based analysis can inform the scientist on which parameters should be considered for revision and which might be left unaltered.

6.2.2 Identification of relevant effects

For illustrative purposes, it is assumed that the highest MMI is computed for parameter θ_1. The scientist chooses to improve the model structure by replacing parameter θ_1 with an opportune state-dependent function. It is not known a-priori how parameter θ_1 should be evolved. Nonetheless, the scientist identifies a set of N_e potentially relevant effects that may be considered in the construction of a function for replacing θ_1. A candidate effect may be a state variable or a combination of state variables (Box and Lucas, 1959). The set of possible effects is denoted as $E = \{\eta_1, ..., \eta_{N_e}\}$, where η_i denotes the i-th effect.

A statistical test is formulated to measure the relevance of the i-th postulated effect for the evolution of parameter θ_1 without the necessity of re-estimating the model parameters. Without loss of generality, the test is formulated to quantify the relevance of effect η_1 on parameter θ_1. The original parameter set is extended by adding an extra parameter $\theta_{N_\theta+1}$. The new parameter set is denoted as $\Theta_e = [\theta_1, ..., \theta_{N_\theta}, \theta_{N_\theta+1}]^T$ (the subscript e stands for effect). The model structure $M_e(M_0 : \theta_1 \rightarrow \eta_1)$ is constructed from the original structure M_0 by replacing parameter θ_1 with the first-order response surface $\theta_1 + \theta_{N_\theta+1} \cdot \eta_1$. The construction of the model structure M_e is illustrated with an example in Figure 6.2. All the model structures $M_e(M_0 : \theta_i \rightarrow \eta_j)\ \forall\, i, j$ are equivalent to the original model structure M_0 under the constraint $\theta_{N_\theta+1} = 0$.

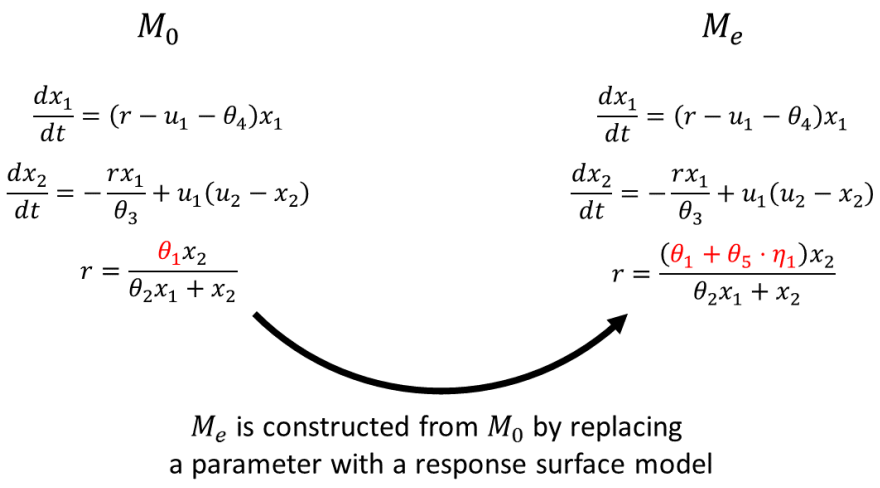


Figure 6.2: Illustrative example of how a model structure M_e is constructed from the original model structure M_0. In this example, the model structure M_e is constructed to assess the relevance of effect η_1 for the evolution of parameter θ_1 into a function. For this purpose, the model structure M_e is constructed from M_0 by replacing parameter θ_1 with the first order response surface $\theta_1 + \theta_5 \cdot \eta_1$.

A Lagrange multipliers test is formulated on $M_e(M_0 : \theta_1 \rightarrow \eta_1)$ to assess the significance of effect η_1 on parameter θ_1. The null and alternative hypothesis considered in the test are formalised mathematically as the presence/absence of a constraint on parameter $\theta_{N_\theta+1}$ as follows

$$\begin{aligned} H_0 : &\quad \theta_{N_\theta+1} = 0 \\ H_a : &\quad \theta_{N_\theta+1} \neq 0 \end{aligned} \tag{6.1}$$

The log-likelihood function is constructed for the model structure M_e with parametrisation Θ_e and it is denoted with the symbol $\mathscr{L}_e(Y|\Theta_e)$. The constrained maximum likelihood estimate $\hat{\Theta}_e$ is obtained by maximising the log-likelihood function $\mathscr{L}_e$ under constraint $\theta_{N_\theta+1} = 0$.

$$\begin{aligned} \hat{\Theta}_e = \arg\max_{\Theta_e} &\mathscr{L}_e(Y|\Theta_e) \\ \text{s.t.} \quad &\theta_{N_\theta+1} = 0 \end{aligned} \tag{6.2}$$

Notice that it is not necessary to solve the optimisation problem in (6.2). In fact, the parameter set $\hat{\Theta}_e = [\hat{\theta}_1, ..., \hat{\theta}_{N_\theta}, 0]^T$ maximises the constrained log-likelihood function and the estimates $\hat{\theta}_1, ..., \hat{\theta}_{N_\theta}$ are already available from the solution of the parameter estimation problem associated with the original model structure M_0.

The parameter set $\hat{\Theta}_e$ also satisfies the constrained maximum likelihood equations

$$\begin{aligned} \nabla\mathscr{L}_e(Y|\hat{\Theta}_e) + \nabla\theta_{N_\theta+1}\hat{\alpha} &= \mathbf{0} \\ \theta_{N_\theta+1} &= 0 \end{aligned} \tag{6.3}$$

where $\hat{\alpha}$ is the value of the Lagrange multiplier associated with the constraint $\theta_{N_\theta+1} = 0$. The computation of the Lagrange multipliers statistic requires the inversion of the $(N_\theta+1)\times(N_\theta+1)$-dimensional Fisher information matrix $\mathbf{H}_e$, which is computed at the constrained maximum likelihood estimate $\hat{\Theta}_e$ using the model structure M_e

$$\mathbf{H}_e = \sum_{i=1}^{N} \nabla\hat{\mathbf{y}}_i(\hat{\Theta}_e)\Sigma_y^{-1}\nabla\hat{\mathbf{y}}_i(\hat{\Theta}_e)^T \tag{6.4}$$

In different words, the computation of the Lagrange multipliers statistic is only possible if the model structure M_e with the extended parameter set Θ_e satisfies the requirements for practical identifiability (for more information on model identifiability see Section 2.5).

In this work, the model M_e will be considered identifiable only if the smallest eigenvalue of $\mathbf{H}_e$ is larger than 1 (Transtrum et al., 2015). If the identifiability requirement is satisfied, the Lagrange multipliers statistic $\xi_e(\eta_1|\theta_1)$ can be computed as in (6.5) and it is asymptotically distributed as a χ^2 with degree of freedom 1 (Aitchison and Silvey, 1958; Silvey, 1959) under H_0.

$$\xi_e(\eta_1|\theta_1) = \hat{\alpha}\nabla\theta_{N_\theta+1}^T \mathbf{H}_e^{-1} \nabla\theta_{N_\theta+1}\hat{\alpha} \sim \chi_1^2 \tag{6.5}$$

In this work, the statistic ξ_e is computed directly as a function of the log-likelihood gradient evaluated at $\Theta_e = \hat{\Theta}_e$ (Rao, 1948) as follows

$$\xi_e(\eta_1|\theta_1) = \nabla\mathscr{L}_e(Y|\hat{\Theta}_e)^T \mathbf{H}_e^{-1} \nabla\mathscr{L}_e(Y|\hat{\Theta}_e) \sim \chi_1^2 \tag{6.6}$$

The statistic $\xi_e(\eta_1|\theta_1)$ was constructed to assess the relevance of effect η_1 for the evolution of parameter θ_1. The procedure can be repeated for all the effects in the set E obtaining the set of statistics $\xi_e(\eta_i|\theta_1)\ \forall\, i = 1,...,N_e$. An Effect Relevance Index (ERI) is proposed as a heuristic measure of relevance of a given effect for the evolution of a given model parameter.

$$\text{ERI}(\eta_i|\theta_1) = \frac{\xi_e(\eta_i|\theta_1)}{\chi_1^2(95\%)} \quad \forall i = 1,...,N_e \tag{6.7}$$

If $\text{ERI}(\eta_i|\theta_1)$ is larger than 1, there is significant evidence to justify the replacement of θ_1 in the model structure M_0 with the response surface $\theta_1 + \theta_5 \cdot \eta_i$. The ERI quantifies the expected improvement in the log-likelihood function as a consequence of an infinitesimal relaxation of the constraint $\theta_{N_\theta+1} = 0$. Hence, if $\text{ERI}(\eta_i|\theta_1) > \text{ERI}(\eta_j|\theta_1) > 1$, one shall expect a more significant improvement in the fitting quality if parameter θ_1 were evolved into the response surface $\theta_1 + \theta_{N_\theta+1} \cdot \eta_i$ rather than the response surface $\theta_1 + \theta_{N_\theta+1} \cdot \eta_j$.

6.2.3 Model evolution

The Effect Relevance Indexes $\text{ERI}(\eta_i|\theta_1)\ \forall i = 1,...,N_e$ quantify the relevance of each effect for the evolution of parameter θ_1. The ERIs provide quantitative information on the expected improvement in the model fitting quality associated with a broad range of modifications in the model structure. The scientist shall proceed by constructing an evolved model structure M_1 from the initial model structure M_0 by replacing parameter θ_1 with an appropriate functional form of the most relevant effect (or effects). Any appropriate functional form may be employed to replace the critical parameter. Nonetheless, in this work, the model

structure M_0 will be evolved by replacing parameters with a first-order response surface of the single most relevant effect. The evolved model structure will be $M_1 \equiv M_e(M_0 : \theta_i \rightarrow \eta_j)$ where θ_i and η_j are respectively the model parameter and the effect associated with the largest computed ERI.

6.3 Case studies

The use of the Effect Relevance Indexes (ERIs) as a support for the evolution of under-fitting model structures is illustrated on two simulated case studies. The case studies represent an extension of the studies presented in Chapter 5 on the diagnosis of model misspecification. In the first case study, presented in Section 6.3.1, the objective is improving the structure of an approximated model of baker's yeast growth by performing an analysis based on the computation of the ERIs. In Section 6.3.2, an ERI-based approach is employed to improve two model structures of glucose-insulin interaction with different level of approximation. The numerical results presented in this section were obtained using Python 3.5 (Python Core Team, 2018).

6.3.1 Case study 1: baker's yeast growth model

6.3.1.1 System model

Baker's yeast growth is assumed to obey the dynamics described by the following system of equations with a Contois-type growth rate

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = (r - u_1 - \theta_4)x_1 \tag{5.10}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \tag{5.11}$$

$$r = \frac{\theta_1 x_2}{\theta_2 x_1 + x_2} \tag{5.12}$$

The system model is the same described in Section 5.3.1.1. The value of the $N_\theta = 4$ parameters in the system model are $\theta^* = [0.310, 0.180, 0.550, 0.050]^T$.

6.3.1.2 Approximated model

The scientist does not know the exact functional form of the system model and proposes an approximated model structure M_0 assuming a Monod-type growth rate

$$M_0 : \begin{cases} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= (r - u_1 - \theta_4)x_1 \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \\ r &= \frac{\theta_1 x_2}{\theta_2 + x_2} \end{cases} \tag{6.8}$$

The approximated model is the same described in Section 5.3.1.2. The identification of the approximated model requires the estimation of $N_\theta = 4$ parameters θ.

6.3.1.3 Objective and methods

The objective in this case study is to improve the structure of the approximated model M_0 described in Section 6.3.1.2 by adopting an ERI-based approach. The same dataset Y used in Section 5.3.1 is used in this case study. The dataset is reported in Appendix F and consists of $N = 28$ samples of $\mathbf{y} = [x_1, x_2]^T$ generated in-silico by integrating the system model and adding Gaussian noise with covariance $\Sigma_y = 2.5 \cdot 10^{-3} \mathbf{I}\ \mathrm{g}^2\,\mathrm{L}^{-2}$. With the considered experimental design and assumed level of system noise, the approximated model structure M_0 is falsified for under-fitting (see Section 5.3.1) with a sum of squared residuals $\chi^2_Y(M_0) = 2210.37$. A MMI-based diagnosis of model misspecification was conducted in Section 5.3.1 showing that the largest model modification index is $\mathrm{MMI}(\theta_2) = 47.08$.

The set of effects $E = \{x_1, x_2, u_1, u_2, x_1^{-1}, x_2^{-1}, u_1^{-1}, u_2^{-1}\}$ is considered for the construction of a function to evolve parameter θ_2. The effect relevance indexes $\mathrm{ERI}(\eta_i|\theta_2)$ $\forall\, i = 1, ..., N_e$ are computed to assess the relevance of all the effects in the set E for the evolution of θ_2 into a state-dependent function. The ERI for a given effect $\eta \in E$ is computed only if the model structure M_e satisfies the requirements for practical identifiability. In this work, the model structure $M_e(M_0 : \theta_2 \to \eta)$ is considered identifiable if the smallest eigenvalue of the Fisher information matrix $\mathbf{H}_e(\hat{\theta}_e)$ is larger than 1 according to White et al. (2016). An evolved model structure M_1 is then constructed by replacing parameter θ_2 in M_0 with the response surface $\theta_2 + \theta_5 \cdot \eta^*$, where η^* is the effect with the highest ERI, i.e. $M_1 \equiv M_e(M_0 : \theta_2 \to \eta^*)$. The model parameters involved in M_1 are estimated by fitting the dataset Y. The fitting of model M_1 is compared with the fitting of model M_0 by performing a Likelihood ratio test (see Section 2.7). The quality of the evolved model structure M_1 is then assessed by performing a two-tailed goodness-of-fit test with 90% of significance. A check on the statistical quality of the parameter estimates is also conducted by performing a 95% t-test.

6.3.1.4 Results

The minimum eigenvalue associated with the model structures $M_e(M_0 : \theta_2 \rightarrow \eta)\ \forall\, \eta \in E$ are reported in Table 6.1. The model structure $M_e(M_0 : \theta_2 \rightarrow x_2)$, obtained by replacing parameter θ_2 in M_0 with the expression $\theta_2 + \theta_5 \cdot x_2$ did not satisfy the requirements for identifiability. Therefore, the ERI$(x_2|\theta_2)$ was not computed. In all the other cases the model structure M_e is identifiable and the ERIs were evaluated. The ERIs associated with the model structure M_0 are reported in Table 6.2. All the computed ERIs are above 1, i.e. ERI$(\eta|\theta_2)\ \forall\, \eta \neq x_2$. Hence, the analysis suggests that any of the considered effects may be relevant for the construction of a function to evolve parameter θ_2. A significant improvement in the model fitting quality is expected should parameter θ_2 be evolved in any response surface $\theta_2 + \theta_5 \cdot \eta$ with $\eta \in E$ s.t. $\eta \neq x_2$. Nevertheless, the highest ERI is ERI$(x_1|\theta_2)$=513.64. Hence, the most significant improvement is expected if θ_2 were replaced with the expression $\theta_2 + \theta_5 \cdot x_1$.

Table 6.1: Baker's yeast system. Computed minimum eigenvalue associated with the evolved models constructed starting from the model structure M_0.

Parameter	Minimum eigenvalue of Information matrix							
to evolve	x_1	x_2	u_1	u_2	x_1^{-1}	x_2^{-1}	u_1^{-1}	u_2^{-1}
θ_2	$1.04 \cdot 10^4$	$1.03 \cdot 10^{-10}$	$1.43 \cdot 10^3$	$1.31 \cdot 10^4$	$3.71 \cdot 10^2$	$1.84 \cdot 10^4$	$1.44 \cdot 10^4$	$4.17 \cdot 10^2$

Table 6.2: Baker's yeast initial model structure M_0. Effect Relevance Indexes associated with the considered effects for the evolution of parameter θ_2. The ERI is not computed and it is not reported (N/A) whenever the model structure M_e does not satisfy the requirements for identifiability.

Parameter	Effect Relevance Indexes							
to evolve	x_1	x_2	u_1	u_2	x_1^{-1}	x_2^{-1}	u_1^{-1}	u_2^{-1}
θ_2	513.64	N/A	162.87	193.83	331.82	104.78	165.22	189.41

An evolved model structure M_1 is constructed by evolving parameter θ_2 into the response surface $\theta_2 + \theta_5 \cdot x_1$, i.e. $M_1 \equiv M_e(M_0 : \theta_2 \rightarrow x_1)$. The model structure M_1 involves the following set of differential and algebraic equations

$$M_1 : \begin{cases} \frac{dx_1}{dt} &= (r - u_1 - \theta_4)x_1 \\ \frac{dx_2}{dt} &= -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \\ r &= \frac{\theta_1 x_2}{\theta_2 + \theta_5 x_1 + x_2} \end{cases} \tag{6.9}$$

The identification of model structure M_1 requires the estimation of a set of $N_\theta = 5$ parameters. Parameters are estimated by fitting the dataset Y. Parameter estimates and related model statistics are reported in Table 6.3. The model M_1 is not falsified by the goodness-of-fit test, i.e. the sum of squared residuals $\chi^2_Y(M_1) = 61.32$ is within the range of acceptability assumed in the test $\chi^2_{N \cdot N_y - N_\theta}(5\%) < \chi^2_Y(M_1) < \chi^2_{N \cdot N_y - N_\theta}(95\%)$. The Likelihood ratio test also suggests that model structure M_1 fits the dataset Y significantly better than the initial model structure M_0

$$\chi^2_Y(M_0) - \chi^2_Y(M_1) = 2149.05 > \chi^2_1(95\%) = 3.84 \tag{6.10}$$

As one can see from Table 6.3, a t-test with 95% of significance highlights that parameter θ_2 in the evolved model structure M_1 is not relevant for fitting the data. A t-value$\ll t_{ref}$ suggests that parameter θ_2 may be constrained to 0 without causing a significant degradation in the model fitting quality. It is observed that under the constraint $\theta_2 = 0$, the model structure M_1 becomes equivalent to the system model structure.

Table 6.3: Baker's yeast model structure M_1. Parameter estimates with related 95% t-values and goodness-of-fit test outcome.

Parameter ID	Parameter estimate	95% t-value* $t_{ref} = 1.68$
θ_1	$3.04 \cdot 10^{-1}$	67.54
θ_2	$1.21 \cdot 10^{-4}$	0.007*
θ_3	$5.37 \cdot 10^{-1}$	43.71
θ_4	$4.54 \cdot 10^{-2}$	11.13
θ_5	$1.82 \cdot 10^{-1}$	20.67
Goodness-of-fit test: Passed		
$\chi^2(5\%)$	χ^2_Y	$\chi^2(95\%)$
35.59	61.32	68.66

*t-value$< t_{ref}$ indicates that a parameter is irrelevant for the fitting.

6.3.2 Case study 2: glucose-insulin interaction model

6.3.2.1 System model

The glucose-insulin regulatory system of a healthy test subject is described by the following set of equations

$$\frac{\mathrm{d}G}{\mathrm{d}t} = -\theta_1(G - G_b) - \theta_2 XG \tag{5.14}$$

$$\frac{\mathrm{d}X}{\mathrm{d}t} = -\theta_3 X + I \tag{5.15}$$

$$\mathrm{IDR} = \max[0, \quad \theta_4(G - \theta_5)t] \tag{5.16}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \mathrm{IDR} - \theta_6 I \tag{5.17}$$

The system model is the same model described in Section 5.3.1.1, where the basal glucose concentration is assumed as $G_b = 93.0$ mg dL^{-1}. The numerical value of the $N_\theta = 6$ parameters appearing in the system model structure is $\theta^* = [2.96 \cdot 10^{-2}, 6.51 \cdot 10^{-6}, 1.86 \cdot 10^{-2}, 5.36 \cdot 10^{-3}, 9.09 \cdot 10^{1}, 2.3 \cdot 10^{-1}]^T$.

6.3.2.2 Approximated models

The scientist does not know the form of the system model and proposes two possible structures to describe the process, namely $M_{0,A}$ and $M_{0,B}$. The model structure $M_{0,A}$ is the following

$$M_{0,A}: \begin{cases} \frac{\mathrm{d}G}{\mathrm{d}t} &= -\theta_1(G - G_b) - \theta_2 X \\ \frac{\mathrm{d}X}{\mathrm{d}t} &= -\theta_3 X + I \\ \mathrm{IDR} &= \max[0, \quad \theta_4(G - \theta_5)t] \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \mathrm{IDR} - \theta_6 I \end{cases} \tag{6.11}$$

and it is the same approximated model structure described in Section 5.3.2.2. The model structure $M_{0,A}$ differs from the structure of the system model in the differential equation describing the glucose concentration in plasma G, where the nonlinear term $\theta_2 XG$ is modelled as a linear term $\theta_2 X$. The identification of the model structure $M_{0,A}$ requires the estimation of $N_\theta = 6$ parameters.

A second model structure $M_{0,B}$ is also considered

$$M_{0,B}: \begin{cases} \frac{dG}{dt} &= -\theta_1 - \theta_2 X \\ \frac{dX}{dt} &= -\theta_3 X + I \\ \text{IDR} &= \max[0, \quad \theta_4(G-\theta_5)t] \\ \frac{dI}{dt} &= \text{IDR} - \theta_6 I \end{cases} \tag{6.12}$$

Also the approximated model structure $M_{0,B}$ differs from the system model structure in the equation describing the concentration of glucose in plasma. In the approximated model structure $M_{0,B}$, the first order derivative of G is expressed as a function of X only. The identification of model structure $M_{0,B}$ also involves the estimation of $N_\theta = 6$ parameters.

6.3.2.3 Objective and methods

The objective in this case study is to diagnose model misspecification with a MMI-based approach and improve the structure of the approximated models using a ERI-based approach. The experimental design adopted in this case study for the generation of dataset Y is the same design adopted in Case A, illustrated in Section 5.3.2.3. The design involves the collection of 23 samples collected in a single IVGTT at initial conditions $G(0) = 298.0$ mg dL^{-1}, $I(0) = 333.0$ μU mL^{-1}, $X(0) = 0.0$ μU min mL^{-1}. Only G and I can be measured, i.e. $\mathbf{y} = [G,I]^T$. Uncorrelated Gaussian system noise is added to the sample with standard deviations 1.0 mg dL^{-1} for measurements of G and 1.5 μU mL^{-1} for measurements of I. The dataset is reported in Table G.1 in Appendix G (only the measurements for G and I in Table G.1 are considered for this case study).

In both model structures $M_{0,A}$ and $M_{0,B}$, parameters are estimated by fitting the dataset Y. Model adequacy is checked with a goodness-of-fit test and in case of under-fitting, model misspecification is diagnosed by computing the MMIs associated with all the model parameters. ERIs are then computed for the parameters with the highest MMI considering the set of possible effects $E = \{G, X, I, G^{-1}\}$ *.

As in the previous case study, the generic ERI$(\eta_j|\theta_i)$ is computed only if the minimum

*The effects X^{-1} and I^{-1} are not considered in this study because the initial value for X is $X(0) = 0.0$ μU min mL^{-1} and I tends to 0.0 μU mL^{-1} in the course of the simulated IVGTT. The numerical integration of the model M_e when the presence of effect X^{-1} is checked is not possible. When the presence of effect I^{-1} is checked, numerical problems in the integration of the system are observed due to a gradient explosion. In some cases, the issue may be solved by performing an algebraic manipulation of the equations in the models M_e. However, algebraic manipulations in the model structures M_e will not be performed in this case study and only models M_e in the form illustrated in Figure 6.2 will be considered.

eigenvalue of the information matrix associated with the structure $M_e(M_0 : \theta_i \rightarrow \eta_j)$ is above 1. Model structures are then evolved following the criterion of the largest computed ERI and parameters are re-estimated by fitting the dataset Y. The improvement of the evolved models compared with the initial model structure is measured by performing a Likelihood ratio test with 95% of significance. The appropriateness of the evolved models in representing the data is quantified by means of a goodness-of-fit test with 95% of significance (see Section 2.7). The statistical quality of the parameter estimates is quantified by means of a t-test with 95% of significance.

6.3.2.4 Results

The parameters in the model structures $M_{0,A}$ and $M_{0,B}$ are estimated by fitting the dataset Y. Estimates for model $M_{0,A}$ are reported in Table 6.4. As one can see, all the model parameters are estimated precisely. All the t-values are above the reference threshold t_{ref}. However, the model is falsified for under-fitting by the two-tailed goodness-of-fit test, i.e., the sum of squared residuals $\chi^2_Y(M_{0,A}) = 97.79$ is above the 95% value $\chi^2(95\%) = 55.75$.

Parameter estimates and related statistics associated with model structure $M_{0,B}$ are reported in Table 6.5. The 95% t-test failed for parameter θ_1 while the other parameters passed the test. This highlights that parameter θ_1 in model structure $M_{0,B}$ may be constrained to 0 without causing a significant loss of fitting quality. Nevertheless, the model structure $M_{0,B}$ is falsified for under-fitting by the goodness-of-fit test. The sum of squared residuals $\chi^2_Y(M_{0,B}) = 220.29$ is above the 95% reference value $\chi^2(95\%) = 55.75$.

A diagnosis of model misspecification based on the MMIs is performed for both model structures. For both models, the largest MMIs are computed for parameters θ_1, θ_2 and θ_3. As one can see from Table 6.4, the MMIs associated with parameters $\theta_1 - \theta_3$ in model structure $M_{0,A}$ are around the value 2.20. The MMIs associated with the model structure $M_{0,B}$ are reported in Table 6.5, where one can see that the MMIs of parameters $\theta_1 - \theta_3$ are around the value 5.38.

The model structures $M_e(M_{0,k} : \theta_i \rightarrow \eta_j)$ with $k = A, B$; $i = 1, 2, 3$ and $j = 1, ..., N_e$ are constructed to assess the relevance of the effects in the set E for the evolution of parameters $\theta_1 - \theta_3$ in both model structures. An identifiability analysis based on the eigendecomposition of the Fisher information matrix is conducted. The minimum eigenvalues are reported for all the model structures M_e constructed starting from model $M_{0,A}$ in Table 6.6 and from model $M_{0,B}$ in Table 6.7. In the $M_{0,A}$ case, the minimum eigenvalue is below 1 in a signifi-

Table 6.4: Glucose-insulin interaction model structure $M_{0,A}$. Parameter estimates with related 95% t-values and MMIs and goodness-of-fit test outcome.

Parameter ID	Parameter estimate	95% t-value* $t_{ref} = 1.68$	MMI
θ_1	$3.44 \cdot 10^{-2}$	45.12	2.20
θ_2	$8.01 \cdot 10^{-4}$	9.10	2.19
θ_3	$2.15 \cdot 10^{-2}$	8.08	2.21
θ_4	$5.33 \cdot 10^{-3}$	18.03	1.17
θ_5	$9.34 \cdot 10^{+1}$	36.38	1.17
θ_6	$2.28 \cdot 10^{-1}$	68.04	1.39
Goodness-of-fit test: Failed for under-fitting			
$\chi^2(5\%)$	χ^2_Y	$\chi^2(95\%)$	
26.50	97.79	55.75	

*t-value$< t_{ref}$ indicates that a parameter is irrelevant for the fitting.

Table 6.5: Glucose-insulin interaction model structure $M_{0,B}$. Parameter estimates with related 95% t-values, MMIs and goodness-of-fit test outcome.

Parameter ID	Parameter estimate	95% t-value* $t_{ref} = 1.68$	MMI
θ_1	$1.93 \cdot 10^{-3}$	0.11*	5.38
θ_2	$1.28 \cdot 10^{-2}$	29.34	5.37
θ_3	$1.73 \cdot 10^{-1}$	22.52	5.39
θ_4	$5.06 \cdot 10^{-3}$	24.70	4.48
θ_5	$9.30 \cdot 10^{+1}$	70.28	3.56
θ_6	$2.26 \cdot 10^{-1}$	74.72	4.58
Goodness-of-fit test: Failed for under-fitting			
$\chi^2(5\%)$	χ^2_Y	$\chi^2(95\%)$	
26.50	220.29	55.75	

*t-value$< t_{ref}$ indicates that a parameter is irrelevant for the fitting.

cant number of cases (see Table 6.6). In the case of model structure $M_{0,B}$, in all the cases, the minumum eigenvalue is below 1 (see Table 6.7). In different words, starting from model structure $M_{0,B}$, whenever a parameter in the range $\theta_1 - \theta_3$ is evolved into a response surface of any of the effects in E the resulting model does not satisfy the requirements set for identifiability with $\theta_e = \hat{\theta}_e$. None of the ERIs is therefore computed for model structure $M_{0,B}$ and the model structure is not evolved.

The ERIs are computed only for the model structure $M_{0,A}$ when the model structure M_e satisfies the requirements for identifiability. The ERIs are reported in Table 6.8. The

Table 6.6: Glucose-insulin interaction system. Computed minimum eigenvalue associated with the model structures $M_e(M_{0,A} : \theta_i \to \eta_j)$ with $i = 1,2,3$ and $j = 1,...,N_e$ constructed starting from the model structure $M_{0,A}$.

Parameter	Minimum Eigenvalue of Information Matrix			
to evolve	G	X	I	G^{-1}
θ_1	$7.21 \cdot 10^{-1}$	3.65	2.11	$2.14 \cdot 10^{-4}$
θ_2	$4.62 \cdot 10^{-1}$	$4.97 \cdot 10^{-2}$	1.85	$3.53 \cdot 10^{-5}$
θ_3	$3.75 \cdot 10^{-2}$	1.48	$4.86 \cdot 10^{-2}$	$1.94 \cdot 10^{-5}$

Table 6.7: Glucose-insulin interaction system. Computed minimum eigenvalue associated with the model structures $M_e(M_{0,B} : \theta_i \to \eta_j)$ with $i = 1,2,3$ and $j = 1,...,N_e$ constructed starting from the model structure $M_{0,B}$.

Parameter	Minimum Eigenvalue of Information matrix			
to evolve	G	X	I	G^{-1}
θ_1	$3.62 \cdot 10^{-5}$	$2.46 \cdot 10^{-6}$	$1.38 \cdot 10^{-2}$	$8.08 \cdot 10^{-9}$
θ_2	$1.40 \cdot 10^{-2}$	$1.41 \cdot 10^{-2}$	$1.26 \cdot 10^{-2}$	$9.13 \cdot 10^{-3}$
θ_3	$1.38 \cdot 10^{-2}$	$1.37 \cdot 10^{-2}$	$1.40 \cdot 10^{-2}$	$1.34 \cdot 10^{-2}$

largest ERI is $\mathrm{ERI}(X|\theta_1) = 60.7$ and it is associated with the evolution of parameter θ_1 into the response surface $\theta_1 + \theta_7 \cdot X$. An evolved model structure $M_{1,A} \equiv M_e(M_{0,A} : \theta_1 \to X)$ is constructed as follows

$$M_{1,A} : \begin{cases} \frac{\mathrm{d}G}{\mathrm{d}t} &= -(\theta_1 + \theta_7 \cdot X)(G - G_b) - \theta_2 X \\ \frac{\mathrm{d}X}{\mathrm{d}t} &= -\theta_3 X + I \\ \mathrm{IDR} &= \max[0, \quad \theta_4(G - \theta_5)t] \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \mathrm{IDR} - \theta_6 I \end{cases} \tag{6.13}$$

Parameters in the model structure $M_{1,A}$ are estimated by fitting dataset Y. The sum of

Table 6.8: Model structure $M_{0,A}$: Effect Relevance Indexes associated with the set of effects $E = \{G, X, I, G^{-1}\}$. The ERI is not computed and it is not reported (N/A) whenever the model structure M_e does not satisfy the requirements for identifiability.

Parameter	Effect Relevance Indexes			
to evolve	G	X	I	G^{-1}
θ_1	N/A	60.7	55.4	N/A
θ_2	N/A	N/A	2.87	N/A
θ_3	N/A	58.5	N/A	N/A

Table 6.9: Generation 2: Glucose-insulin interaction model. Parameter estimates with related 95% t-values and goodness-of-fit test outcome.

Parameter ID	Parameter estimate	95% t-value* $t_{ref} = 1.68$
θ_1	$2.69 \cdot 10^{-2}$	12.73
θ_2	$5.94 \cdot 10^{-4}$	5.95
θ_3	$1.74 \cdot 10^{-2}$	4.95
θ_4	$5.30 \cdot 10^{-3}$	19.39
θ_5	$9.12 \cdot 10^{+1}$	40.23
θ_6	$2.29 \cdot 10^{-1}$	67.89
θ_7	$8.70 \cdot 10^{-6}$	4.11
Goodness-of-fit test: Passed		
$\chi^2(5\%)$	χ^2_Y	$\chi^2(95\%)$
25.69	39.09	54.57

*t-value$< t_{ref}$ indicates that a parameter is irrelevant for the fitting.

squared residuals associated with the evolved model $M_{1,A}$ is $\chi^2_Y(M_{1,A}) = 39.09$. Parameter estimates and related t-values are reported in Table 6.9. All the parameters pass the 95% t-test, i.e., the t-value is above t_{ref} for all parameters. This suggests that fixing any parameter to 0 is expected to result in a significant degradation of the fitting quality. In particular, the introduced parameter θ_7 associated with the presence of effect X is detected as relevant. The improvement achieved by model structure $M_{1,A}$ compared with the initial model $M_{0,A}$ is also demonstrated by the failed Likelihood ratio test

$$\chi^2_Y(M_{0,A}) - \chi^2_Y(M_{1,A}) = 58.7 > \chi^2_1(95\%) = 3.84 \tag{6.14}$$

The failed Likelihood ratio test suggests that model structure $M_{1,A}$ should be preferred over the initial model structure $M_{0,A}$. The evolved model structure $M_{1,A}$ passes the goodness-of-fit test , i.e., $\chi^2(5\%) < \chi^2_Y(M_{1,A}) < \chi^2(95\%)$.

6.3.3 Results discussion

Two simulated case studies were presented in this Chapter to demonstrate the use of MMIs and ERIs as tools for improving the structure of approximated kinetic models.

In Case study 1, a baker's yeast bioreactor system was considered. The objective was to improve an approximated model in the form of the system of equations M_0 in (6.8). A model misspecification diagnosis based on the computation of the MMIs highlighted that a significant improvement in the model fitting quality was expected should parameter θ_2 be

evolved into a function of the state variables. The largest ERI associated with parameter θ_2 was $\mathrm{ERI}(x_1|\theta_2)$, suggesting that the most relevant effect for the evolution of θ_2 is the state x_1, namely the concentration of biomass in the bioreactor. Parameter θ_2 appears at the denominator of the rate expression. Hence, it was possible to detect the presence of an inhibiting effect of biomass concentration on the growth rate that was not considered in the initially available model M_0. An evolved model M_1 (6.9) was constructed by evolving parameter θ_2 into the response surface $\theta_2 + \theta_5 \cdot x_1$, which included the main detected effect. The evolved model structure M_1 achieved a significantly better fitting compared with the initial structure M_0 and was not falsified by the goodness-of-fit test. The evolved model structure M_1 is indistinguishable from the system model structure described in Section 5.3.1.1, which was used to generate the in-silico dataset Y. In different words, it is not possible to design an experiment with the aim of discriminating between the model structure M_1 and the system model.

In Case Study 2, the glucose-insulin regulatory system of a healthy test subject was considered. The objective was to improve the structure of two approximated kinetic models, namely model $M_{0,A}$ (6.11) and model $M_{0,B}$ (6.12). Both model structures differ from the system model in the form of the equation describing the glucose concentration in plasma G. In fact, while the system model equation includes a linear effect of G and a nonlinear effect XG, the approximated model $M_{0,A}$ only includes a linear effect for G and for X; the model structure $M_{0,B}$ includes only a linear effect of X on the concentration of glucose in plasma. A MMI-based analysis highlighted that, in both model $M_{0,A}$ and model $M_{0,B}$, the model parameters θ_1, θ_2 and θ_3 are those that are expected to reduce process-model mismatch the most should any of them be evolved in a state-dependent function.

Starting from model structure $M_{0,A}$, the set of models $M_e(M_{0,A} : \theta_i \rightarrow \eta)$ with $i = 1,2,3$ and $\eta \in E$ was constructed. The minimum eigenvalue computed for the information matrices associated with these model structures was always below 4.27 and in a significant number of cases the minimum eigenvalue was below 1 (see Table 6.6). In such cases, the ERI was not evaluated. The largest ERI was computed for effect X on parameter θ_1, namely $\mathrm{ERI}(X|\theta_1)$. The evolved model structure $M_{1,A}$, constructed by replacing parameter θ_1 in $M_{0,A}$ with the response surface $\theta_1 + \theta_7 \cdot X$ was not falsified by the goodness-of-fit test.

It is recognised that such model structure is equivalent to the system model structure, in fact

$$\begin{aligned} \frac{\mathrm{d}G}{\mathrm{d}t} &= -(\theta_1 + \theta_7 X)(G - G_b) - \theta_2 X \\ &= -\theta_1 (G - G_b) - \theta_7 X G + (\theta_7 G_b - \theta_2) X \end{aligned} \tag{6.15}$$

The term $(\theta_7 G_b - \theta_2)X$ is not present in the system model structure. The discrepancy between model $M_{1,A}$ and system model vanishes if the constraint $\mathbf{s} = [\theta_7 G_b - \theta_2] = \mathbf{0}$ is enforced in the parameter estimation problem. This is validated through a Wald test (see Section 2.7). The statistic is computed as

$$\xi_W = \mathbf{s}(\hat{\theta})^T \left[\nabla \mathbf{s}^T \mathbf{V}_\theta \nabla \mathbf{s}\right]^{-1} \mathbf{s}(\hat{\theta}) = 3.34 \tag{6.16}$$

Since $\xi_W < \chi^2_1(95\%) = 3.84$ it is concluded that there is no evidence to disprove the presence of the constraint.

It is also recognised that the model structure $M_e(M_{0,A} : \theta_2 \to G)$ is equivalent to the system model structure under the constraint $\theta_2 = 0$. However, the model structure $M_e(M_{0,A} : \theta_2 \to G)$ did not satisfy the requirement for identifiability at the constrained estimate $\hat{\theta}_e$ and the associated effect relevance index $\mathrm{ERI}(X|\theta_2)$ was not computed. None of the model structures M_e constructed from model $M_{0,B}$, i.e., $M_e(M_{0,B} : \theta \to \eta)$, satisfied the requirements for identifiability and no ERI was computed in this case.

6.3.4 Computational times and problem size

In the baker's yeast case study, the model M_0 is the same approximated model considered in Section 5.3.1.2 and involves $N_\theta = 4$ parameters. Also, the dataset Y used for the computation of the ERIs is the same used in Section 5.3.1 and consisted of $N = 28$ samples where each sample involved $N_y = 2$ measured quantities, namely the biomass concentration x_1 and the substrate concentration x_2. The computation of the ERIs associated with the approximated model M_0 required around 10.0 s of CPU time. The times required for the evaluation of the ERIs are reported in Table 6.10.

In the case study on the glucose-insulin regulatory system, the model structure $M_{0,A}$ is the same approximated model considered in Section 5.3.2.2, which involved 3 ODEs and included $N_\theta = 6$ parameters. The dataset Y is also the same considered in Section 5.3.2.3 and consisted of $N = 23$ samples, each involving $N_y = 2$ measured quantities, namely the glucose concentration in plasma G and the insulin concentration in plasma I. The computational times required for the evaluation of each ERI is around 18.0 s, as shown in Table

6.11.

The higher computational time in the glucose-insulin case is associated primarily with the increase in the number of parameters. The number of parameters affects the dimension of the log-likelihood gradient and the dimension of the Fisher information matrix. A higher number of parameters results in an increase in the computational cost associated with the evaluation of the sensitivities, and the computation and inversion of the Fisher information matrix. However, the main advantage of the ERI-based approach is that the model parameters are not re-estimated to assess the relevance of the effects and therefore, the possible numerical failures associated with the re-estimation of parameters are avoided.

Table 6.10: Baker's yeast system. Computational time expressed in [s] required for the computation of the ERIs.

Parameter	Computational time for ERI evaluation [s]							
to evolve	x_1	x_2	u_1	u_2	x_1^{-1}	x_2^{-1}	u_1^{-1}	u_2^{-1}
θ_2	10.42	N/A	10.81	10.80	10.38	10.41	10.78	10.78

Table 6.11: Glucose-insulin regulatory system. Computational times associated with the computation of the ERIs of model structure $M_{0,A}$.

Parameter	Computational time for ERI evaluation [s]			
to evolve	G	X	I	G^{-1}
θ_1	N/A	17.88	17.87	N/A
θ_2	N/A	N/A	17.92	N/A
θ_3	N/A	18.05	N/A	N/A

6.4 Limitations of the ERI-based approach

In the computation of the ERIs, the Fisher information matrix $\mathbf{H}_e(\hat{\theta}_e)$ is evaluated at the constrained maximum likelihood estimate $\hat{\theta}_e$, i.e., under the constraint that $\theta_{N_\theta+1} = 0$. The Lagrange multipliers test assumes that the matrix $\mathbf{H}_e(\hat{\theta}_e)$ well approximates the information matrix at the unconstrained maximum likelihood estimate (Silvey, 1959). This assumption may not hold whenever the model is nonlinear in the parameters and the unconstrained estimate is *far* from the constrained estimate (Silvey, 1959; Buse, 1982). It is therefore possible that the information matrix evaluated at the constrained estimate is singular or nearly singular when the actual information matrix evaluated at the unconstrained estimate is well-conditioned.

In the glucose-insulin case study, most models M_e did not satisfy the identifiability requirement at the constrained maximum likelihood estimate (see Section 6.3.2.4). In all these cases, an ERI-based approach cannot be applied and a re-estimation of the parameters would be required to check for the relevance of a given effect. Whenever the matrix $\mathbf{H}_e(\hat{\theta}_e)$ is singular or nearly singular, one may proceed by estimating the unconstrained maximum likelihood estimate and test model identifiability after the re-estimation of the parameters. Nonetheless, also in this case identifiability problems may hamper the application of an ERI-based approach. This is recognised as an important limitation of the proposed methodology. In fact, it may happen that a transition through a non-identifiable model structure is required to evolve towards an appropriate model that is not falsified by the observations.

It is important to observe that the ERI is a local measure of model improvement and misleading results may be obtained from an ERI-based analysis also when the information matrix $\mathbf{H}_e(\hat{\theta}_e)$ is invertible. A qualitative example is reported in Figure 6.3, which shows the log-likelihood profiles associated with two superstructures including respectively effect η_A (dark line) and effect η_B (red line). As one can see, the peak in the log-likelihood associated with effect η_A is higher than the peak associated with the effect η_B. However, the ERI for both models is evaluated at the constrained estimate, i.e., under the constraint $\theta_{N_\theta+1} = 0$, where the gradient associated with effect η_B is steeper than the gradient associated with η_A. If the model is nonlinear in the parameters, the Fisher information matrix evaluated at the constrained estimate may not be representative of the actual curvature at the peak of the log-likelihood. Hence, the Lagrange multipliers statistic may fail to accurately quantify how far is the unconstrained estimate from the constrained estimate. It is therefore possible that the expected rate of improvement in the log-likelihood measured at the constrained estimate is higher for η_B than for effect η_A. This would lead to the misleading conclusion that effect η_B is more relevant than η_A for the improvement of the model fitting quality.

6.5 Final remarks

A procedure for supporting the scientist in the improvement of models in the presence of significant process-model mismatch was illustrated in this Chapter. The procedure follows from the assumption that the improvement of a model structure requires the evolution of a certain model parameter into an opportune state-dependent function. The selection of relevant parameters that should be considered for evolution may follow from an analysis based on the computation of the MMIs or from the modeller's insight on the physical system.

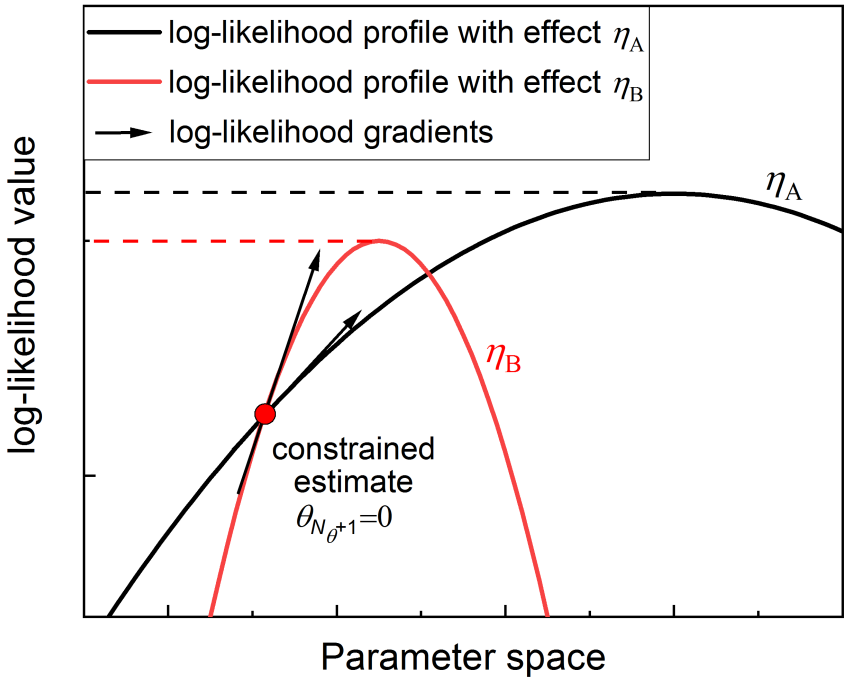


Figure 6.3: Qualitative diagram showing the locality of the ERI-based approach for the detection of relevant effects. In the figure, two different effects η_A ad η_B are tested. The maximum likelihood estimate associated with effect η_A is higher than the maximum likelihood estimate achievable with the inclusion of η_B in the model. However, the relevance of the effects in a ERI-based framework is checked only locally at the constrained maximum likelihood estimate. Assuming that the information associated with the two effects is the same, an analysis based on the ERIs would suggest that effect η_B is more relevant than η_A. In fact, the rate of change of the log-likelihood at the constrained estimate is higher in the η_B case than in the η_A case.

Once relevant parameters are selected for evolution, the modeller proposes a set of effects that may be relevant for the evolution of those parameters.

The procedure continues with the construction of a set of superstructures so that *i*) each superstructure includes a different effect compared with the initial approximated model *ii*) the presence of the extra effect in each superstructure is controlled by an additional parameter $\theta_{N_\theta+1}$ *iii*) each superstructure is equivalent to the original model under the constraint $\theta_{N_\theta+1} = 0$. A Lagrange multipliers test is then performed on each superstructure with the aim of disproving the presence of the constraint.

An Effect Relevance Index (ERI) is proposed in this Chapter as a normalised Lagrange multipliers statistic to quantify the relevance of each postulated effects for the evolution of the approximated model without re-estimating the model parameters. The ERI provides a quantification for the expected rate of change of the log-likelihood profile at the constrained maximum likelihood estimate. The inclusion in the model of effects with high ERI is ex-

pected to produce a more significant improvement on the fitting than the inclusion of effects with low ERI. The ERI-based approach was demonstrated on simulated case studies where the objective was to improve the structure of three approximated models, namely a model of baker's yeast growth M_0 and two models of glucose-insulin interaction $M_{0,A}$ and $M_{0,B}$. The ERI-based approach led to a significant improvement of the approximated model structures M_0 and $M_{0,A}$ and the identification of models that were equivalent to their respective system models. In the case of model structure $M_{0,B}$, however, the ERIs were not computed because their associated information matrix was nearly singular (i.e., the minimum eigenvalue of the information matrix was smaller than 1).

The limitations of the ERI-based approach for effect detection are associated primarily with the locality of the Lagrange multipliers test and with the fact that the information matrix may be extremely sensitive to a change in the model parameters when the model is nonlinear in the parameters. As a consequence, it may be not be possible to accurately quantify the model improvement associated with the inclusion of a given effect by performing an ERI-based analysis at the constrained estimate. In such conditions, a re-estimation of the parameters in the presence of the postulated effects may be required. When both the constrained and the unconstrained estimates are available for each effect, effect relevance may be quantified by employing a likelihood ratio test (Wilks, 1938), the Akaike Information criterion (Akaike, 1974) or the Bayesian information criterion (Schwarz, 1978). In future works, further frameworks for model improvement will be tested including a step of re-estimation of the model parameters and advocate the computation of more accurate indexes to quantify effect relevance.

Chapter 7

Conclusion and future perspectives

The identification of a kinetic model requires *1*) the formulation of an appropriate model structure *2*) the estimation of its kinetic parameters by fitting experimental data and *3*) the validation of the model predictions against experimental observations. Thanks to the invaluable contribution of many scientists in the fields of automation and model building, automated platforms for kinetic model identification are now becoming a reality. Nonetheless, it is argued that there are still a number of computational challenges that need to be addressed to promote the diffusion of these platforms in research laboratories.

These challenges are associated with aspects of the modelling activity that cannot be effectively automated with current model building techniques. Some of these aspects are *i*) the definition of an appropriate set of modelling assumptions and their translation into a set of model equations *ii*) the estimation of parameters and the optimal MBDoE for parameter precision in the presence of approximated model structures *iii*) the improvement of approximated model structures embracing the available experimental evidence and *iv*) the robust estimation of parameters in the presence of model sloppiness. The aim of this research project is the formulation of robust modelling frameworks to systematically address these challenges.

A number of novel techniques for model identification and refinement are proposed in this Thesis to make the kinetic modelling activity more systematic and less sensitive to human error and bias. The main scientific contributions presented in this Thesis are:

1. An online Reparametrisation (RP) method to automatically reduce the chance of numerical failures associated with model sloppiness in the course of online kinetic modelling studies.

2. A systematic framework for the online identification of kinetic models in the presence

of structural model uncertainty. In the frameworks, model parameters are inferred together with the geometry of the model reliability domain, which is returned by the algorithm in the form of a model reliability map. A conservative MBDoE criterion for the identification of approximated models is proposed where the research of optimal experimental conditions is constrained within the domain of model reliability.

3. A Model Modification Index (MMI) based on maximum likelihood inference for diagnosing model misspecification in under-fitting models, i.e., in the presence of a significant process-model mismatch.

4. An Effect Relevance Index (ERI) to rapidly evaluate effective strategies to improve the structure of an approximated model when under-fitting is detected.

Online-RP was tested both in-silico and in an automated platform for the identification of a 2-parameter model of benzoic acid esterification with ethanol in a microreactor. It was shown that the application of online-RP led to the minimisation of the condition number associated with the parameter estimation problem in 4 iterations, i.e., after the collection of 4 samples from the initial algorithm call. This resulted in a more robust estimation of the model parameters compared with a standard model identification algorithm. It was also shown that the computational burden associated with the identification of the model is not significantly affected by the introduction of the online-RP step. Future work on the online-RP framework shall focus primarily on three aspects: *i*) improving the initialisation of the algorithm to reduce the number of iterations required to bring the condition number to unity; *ii*) validating the framework on more complex model structures, e.g., in the presence of a higher number of parameters and/or measured system states; *iii*) extending the framework to include also nonlinear transformations of the parameter space.

The proposed framework for the identification of models under structural model uncertainty was applied on two case studies *1*) a simulated case where it was applied online on the identification of an approximated model of ethanol dehydrogenation on copper-based catalyst *2*) a real case where it was employed offline to identify an approximated model of methanol oxidation on silver catalyst. In both cases, it was shown that it is possible to effectively identify approximated models by using only the data collected within the model reliability domain. The model reliability maps returned by the algorithm may be employed to quantify the expected model accuracy in unexplored experimental conditions and assess

whether the available model structure is appropriate for a specific task or whether a change in the model structure is required. Future research activities shall aim at implementing the proposed framework in an automated kinetic modelling platform and validate the approach on the online identification of kinetic models. The proposed framework may also offer a basis for the formulation of alternative criteria for model selection based on reliability maps. As an example, the problem of choosing the *best* model among a set of candidates may be recast in terms of selecting the model with the largest domain of reliability.

If a modification in the model structure is required, a MMI-based model diagnosis may be employed to detect which parameters in the model are most likely to hide state dependencies. Parameters with the highest MMI are those that are expected to improve the model fitting quality the most should they be evolved into state-dependent expressions. A ERI-based approach may then be employed to detect which effects are the most relevant for the construction of opportune functions to replace model parameters. The approach was tested on simulated case studies where the aim was to improve the structure of an approximated model of baker's yeast growth in a bioreactor and approximated models of the glucose-insulin regulatory system of a healthy test subject. The MMI and ERI represent computationally inexpensive heuristics. In fact, their computation does not require a re-estimation of the model parameters. Nonetheless, it is recognised that both heuristics only represent a local measure of model improvement. In particular, it was shown that a ERI-based analysis may lead to an inaccurate quantification of the relevance of a given effect if the model is nonlinear in the parameters. Furthermore, the computation of both MMIs and ERIs requires the inversion of information matrices. Unless an appropriate experimental design is employed for the collection of the dataset, such information matrices may not be invertible. In addition to the previous aspects, the MMI proposed in Chapter 5 was formulated neglecting parameter interaction. A possible formulation of a multivariate MMI is reported in Appendix H, where some necessary conditions for its computation are derived. Future research activities shall aim at *i*) identifying sufficient conditions for the computation of a multivariate MMI to account for parameter interaction in the diagnosis and *ii*) formulating and validating experimental design approaches to handle cases where the information matrix is not invertible and make the computation of MMIs and ERIs feasible.

The integration of the aforementioned approaches may also offer the basis for interesting future research developments. It is conjectured that a combined application of online-

RP with the MBDM estimator may represent a powerful tool for online kinetic modelling studies in the presence of both model sloppiness and approximated modelling assumptions. A further research direction may focus on the employment of MMI-based and ERI-based approaches for improving the reliability of an approximated model specifically within the domain of application, i.e., the range of experimental conditions where the modeller wants the model to be accurate. The proposed RP approach may also be employed (either offline or online), on the extended parameter spaces associated with the computation of MMIs and ERIs. It is speculated that the application of a RP approach may contribute to reducing the chance of numerical issues that may occur in the computation of Lagrange multipliers statistics when the condition number of the information matrix is high.

It is expected that the employment of the proposed modelling frameworks may be particularly beneficial in the identification and improvement of models for complex biological systems, e.g. models of algae growth for biofuel production (Zhang et al., 2015) and models of biomass conversion (Ranzi et al., 2008). In fact, systems in bioengineering are recognised to be extremely challenging to model due to a high requirement for time and resources in the experimentation, model sloppiness and poorly understood dynamics. It is also in the aims of future research to validate the proposed modelling frameworks on case studies outside the field of process systems engineering, particularly in the areas of haematology, physiology and pharmacology. In fact, the identification of accurate kinetic models in all the aforementioned areas relies on an efficient extraction of information from small datasets in order to reduce the distress caused to test subjects. In the field of haematology, the proposed modelling frameworks may be employed to improve fundamental understanding on the mechanisms of tumour growth *in-vivo*, advocating a more rapid quantification of the potency of specific cancer treatments (Klinke and Wang, 2017). The proposed modelling frameworks could be also employed for better understanding human physiology, enabling the development of robust model-based diagnostic tools for healthcare applications (Galvanin et al., 2014). In pharmacology, modelling algorithms sprouting from this research project could be validated on the identification of pharmacodynamic models to describe the dynamic response of bacterial species to specific pharmaceutical treatments *in-vitro* (Foerster et al., 2016). A further validation of the proposed tools for diagnosis and improvement of approximated models could be conducted on the identification of pharmacokinetic models to understand and quantify the response of test subjects to given clinical

protocols, advocating the design of more effective, safer and less invasive clinical trials (Abbiati et al., 2018).

In the frameworks illustrated in this Thesis, some prior knowledge on the system dynamics is provided as an input to the model identification algorithm in the form of a candidate model structure (possibly approximated). Future work should focus on developing cognitive algorithms for model identification that do not require such prior knowledge. Recent advances in the fields of artificial intelligence, reinforcement learning and genetic programming suggest the feasibility of constructing surrogate cognitive agents (SCAs) in the form of algorithms executed in computational frameworks (Mnih et al., 2015). The recent application of artificial neural networks and genetic algorithms to complex control and design problems led to solutions that were previously thought to be achievable only by a cognitive agent. It was also demonstrated that SCAs may even surpass cognitive human-level capabilities in achieving pre-defined goals in pre-defined environments (Mnih et al., 2015; Chen et al., 2016; Nourbakhsh et al., 2016). Currently, SCAs are being trained either on insufficient data or on virtual realities that have little connection with the physical world, e.g. chessboards (Campbell et al., 2002) or Atari videogames (Mnih et al., 2015). Nevertheless, the technology required to build a surrogate scientist that can autonomously learn how to build kinetic model structures embracing first principles, identifiability constraints and experimental evidence may soon be available.

Bibliography

Abbiati, R. A., Savoca, A., and Manca, D. (2018). Chapter 2 - An engineering oriented approach to physiologically based pharmacokinetic and pharmacodynamic modeling. In Manca, D., editor, *Computer Aided Chemical Engineering*, volume 42 of *Quantitative Systems Pharmacology*, pages 37–63. Elsevier.

Adjiman, C. S., Schweiger, C. A., and Floudas, C. A. (1998). Mixed-Integer Nonlinear Optimization in Process Synthesis. In Du, D.-Z. and Pardalos, P. M., editors, *Handbook of Combinatorial Optimization: Volume1–3*, pages 1–76. Springer US, Boston, MA.

Agarwal, A. K. and Brisk, M. L. (1985a). Sequential experimental design for precise parameter estimation. 1. Use of reparameterization. *Industrial & Engineering Chemistry Process Design and Development*, 24(1):203–207.

Agarwal, A. K. and Brisk, M. L. (1985b). Sequential experimental design for precise parameter estimation. 2. Design criteria. *Industrial & Engineering Chemistry Process Design and Development*, 24(1):207–210.

Aitchison, J. and Silvey, S. D. (1958). Maximum-Likelihood Estimation of Parameters Subject to Restraints. *The Annals of Mathematical Statistics*, 29(3):813–828.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 199–213. Springer, New York, NY.

Alberton, A. L., Schwaab, M., Schmal, M., and Pinto, J. C. (2009). Experimental errors in kinetic tests and its influence on the precision of estimated parameters. Part I—Analysis of first-order reactions. *Chemical Engineering Journal*, 155(3):816–823.

Alshraideh, H. and Runger, G. (2014). Process Monitoring Using Hidden Markov Models. *Quality and Reliability Engineering International*, 30(8):1379–1387.

Amrhein, M., Srinivasan, B., and Bonvin, D. (1999). Target factor analysis of reaction data: use of data pre-treatment and reaction-invariant relationships. *Chemical Engineering Science*, 54(5):579–591.

Andreasen, A., Lynggaard, H., Stegelmann, C., and Stoltze, P. (2003). A microkinetic model of the methanol oxidation over silver. *Surface Science*, 544(1):5–23.

Andreasen, A., Lynggaard, H., Stegelmann, C., and Stoltze, P. (2005). Simplified kinetic models of methanol oxidation on silver. *Applied Catalysis A: General*, 289(2):267–273.

Anguelova, M., Karlsson, J., and Jirstrand, M. (2012). Minimal output sets for identifiability. *Mathematical Biosciences*, 239(1):139–153.

Anselin, L. (1988). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, 20(1):1–17.

Asprey, S. and Macchietto, S. (2002). Designing robust optimal dynamic experiments. *Journal of Process Control*, 12:545–556.

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

Asprey, S. P. and Naka, Y. (1999). Mathematical Problems in Fitting Kinetic Models - Some New Perspectives. *Journal of Chemical Engineering of Japan*, 32(3):328–337.

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

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

Audoly, S., Bellu, G., D'Angiò, L., Saccomani, M. P., and Cobelli, C. (2001). Global identifiability of nonlinear models of biological systems. *IEEE transactions on bio-medical engineering*, 48(1):55–65.

Bah, B. (2008). Diffusion Maps: Analysis and Applications. Master's thesis, University of Oxford.

Banzhaf, W., Nordin, P., Keller, R. E., and Francone, F. D. (2015). *Genetic Programming: An Introduction*. Morgan Kaufmann, Burlington, MA.

Barber, D. (2011). *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge ; New York.

Bard, Y. (1974). *Nonlinear Parameter Estimation*. Academic Press, New York.

Bardow, A. (2008). Optimal experimental design of ill-posed problems: The METER approach. *Computers & Chemical Engineering*, 32(1):115–124.

Barz, T., López Cárdenas, D. C., Arellano-Garcia, H., and Wozny, G. (2013). Experimental evaluation of an approach to online redesign of experiments for parameter determination. *AIChE Journal*, 59(6):1981–1995.

Barz, T., López Cárdenas, D. C., Cruz Bournazou, M. N., Körkel, S., and Walter, S. F. (2016). Real-time adaptive input design for the determination of competitive adsorption isotherms in liquid chromatography. *Computers & Chemical Engineering*, 94:104–116.

Bassett, R. and Deride, J. (2019). Maximum a posteriori estimators as a limit of Bayes estimators. *Mathematical Programming*, 174(1):129–144.

Bellu, G., Saccomani, M. P., Audoly, S., and D'Angiò, L. (2007). DAISY: a new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, 88(1):52–61.

Benabbas, L., Asprey, S. P., and Macchietto, S. (2005). Curvature-Based Methods for Designing Optimally Informative Experiments in Multiresponse Nonlinear Dynamic Situations. *Industrial & Engineering Chemistry Research*, 44(18):7120–7131.

Benson, S. W. and Buss, J. H. (1958). Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *The Journal of Chemical Physics*, 29(3):546–572.

Bera, A. K. and Bilias, Y. (2001). Rao's score, Neyman's C(α) and Silvey's LM tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference*, 97(1):9–44.

Bergman, R. N., Ider, Y. Z., Bowden, C. R., and Cobelli, C. (1979). Quantitative estimation of insulin sensitivity. *The American Journal of Physiology*, 236(6):E667–677.

Bergman, R. N., Phillips, L. S., and Cobelli, C. (1981). Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose. *The Journal of Clinical Investigation*, 68(6):1456–1467.

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Bhan, A., Hsu, S.-H., Blau, G., Caruthers, J. M., Venkatasubramanian, V., and Delgass, W. N. (2005). Microkinetic modeling of propane aromatization over HZSM-5. *Journal of Catalysis*, 235(1):35–51.

Bhattacharjee, B. (2003). *Kinetic model reduction using integer and semi-infinite programming*. Thesis, Massachusetts Institute of Technology.

Biegler, L. T., Grossmann, I. E., and Westerberg, A. W. (1997). *Systematic Methods of Chemical Process Design*. Prentice Hall, Upper Saddle River, N.J.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA.

Bonvin, D., Georgakis, C., Pantelides, C. C., Barolo, M., Grover, M. A., Rodrigues, D., Schneider, R., and Dochain, D. (2016). Linking Models and Experiments. *Industrial & Engineering Chemistry Research*, 55(25):6891–6903.

Bonvin, D. and Rippin, D. W. T. (1990). Target factor analysis for the identification of stoichiometric models. *Chemical Engineering Science*, 45(12):3417–3426.

Bournazou, M. N. C., Barz, T., Nickel, D. B., López Cárdenas, D. C., Glauche, F., Knepper, A., and Neubauer, P. (2016). Online optimal experimental re-design in robotic parallel fed-batch cultivation facilities. *Biotechnology and Bioengineering*, 114(3):610–619.

Box, G. E. and Hill, W. J. (1967). Discrimination Among Mechanistic Models. *Technometrics*, 9(1):57–71.

Box, G. E. P. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley.

Box, G. E. P. and Lucas, H. L. (1959). Design of Experiments in Non-Linear Situations. *Biometrika*, 46(1/2):77–90.

Box, G. E. P. and Wilson, K. B. (1951). On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(1):1–45.

Brenan, K. E., Campbell, S. L., and Petzold, L. R. (1987). *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition.

Brendel, M. and Marquardt, W. (2008). Experimental design for the identification of hybrid reaction models from transient data. *Chemical Engineering Journal*, 141(1):264–277.

Breusch, T. S. and Pagan, A. R. (1980). The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. *The Review of Economic Studies*, 47(1):239–253.

Broadbelt, L. J., Stark, S. M., and Klein, M. T. (1994). Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates. *Industrial & Engineering Chemistry Research*, 33(4):790–799.

Bruwer, M.-J. and MacGregor, J. F. (2006). Robust multi-variable identification: Optimal experimental design with constraints. *Journal of Process Control*, 16(6):581–600.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, 2nd edition.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304.

Buse, A. (1982). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *The American Statistician*, 36(3):153–157.

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

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

Buzzi-Ferraris, G., Forzatti, P., and Paolo, C. (1990). An improved version of a sequential design criterion for discriminating among rival multiresponse models. *Chemical Engineering Science*, 45(2):477–481.

Buzzi-Ferraris, G. and Manenti, F. (2009). Kinetic models analysis. *Chemical Engineering Science*, 64(5):1061–1074.

Campbell, M., Hoane, A. J., and Hsu, F.-h. (2002). Deep Blue. *Artificial Intelligence*, 134(1):57–83.

Cao, E. and Gavriilidis, A. (2005). Oxidative dehydrogenation of methanol in a microstructured reactor. *Catalysis Today*, 110(1–2):154–163.

Carotenuto, G., Tesser, R., Di Serio, M., and Santacesaria, E. (2013). Kinetic study of ethanol dehydrogenation to ethyl acetate promoted by a copper/copper-chromite based catalyst. *Catalysis Today*, 203:202–210.

Chakrabarty, A., Buzzard, G. T., and Rundell, A. E. (2013). Model-based design of experiments for cellular processes. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 5(2):181–203.

Chambers, E. A. and Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika*, 54(3-4):573–578.

Chandra, T. K. and Joshi, S. N. (1983). Comparison of the Likelihood Ratio, Rao's and Wald's Tests and a Conjecture of C. R. Rao. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 45(2):226–246.

Chen, Y., Elenee Argentinis, J., and Weber, G. (2016). IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clinical Therapeutics*, 38(4):688–701.

Chiavazzo, E., Gear, C. W., Dsilva, C. J., Rabin, N., and Kevrekidis, I. G. (2014). Reduced Models in Chemical Kinetics via Nonlinear Data-Mining. *Processes*, 2(1):112–140.

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

Chiş, O. T., Banga, J. R., and Balsa-Canto, E. (2014). Sloppy models can be identifiable. *arXiv:1403.1417 [q-bio]*.

Chou, C. P. and Bentler, P. M. (1990). Model Modification in Covariance Structure Modeling: A Comparison among Likelihood Ratio, Lagrange Multiplier, and Wald Tests. *Multivariate Behavioral Research*, 25(1):115–136.

Collins, J. R. (1982). Robust M-estimators of location vectors. *Journal of Multivariate Analysis*, 12(4):480–492.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, page 23.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24(2):406–424.

Cozad, A., Sahinidis, N. V., and Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6):2211–2227.

Cozad, A., Sahinidis, N. V., and Miller, D. C. (2015). A combined first-principles and data-driven approach to model building. *Computers & Chemical Engineering*, 73:116–127.

Cubillos, F. A., Acuña, G., and Lima, E. L. (2007). Real-time process optimization based on grey-box neural models. *Brazilian Journal of Chemical Engineering*, 24(3):433–443.

Del Rio Chanona, E. A., Graciano, J. E. A., Bradford, E., and Chachuat, B. (2019). Modifier-Adaptation Schemes Employing Gaussian Processes and Trust Regions for Real-Time Optimization. *IFAC-PapersOnLine*, 52(1):52–57.

Devore, J. L. (2010). *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, Inc, Boston, MA, 8th edition.

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

Diwekar, U. M. and Rubin, E. S. (1991). Stochastic modeling of chemical processes. *Computers & Chemical Engineering*, 15(2):105–114.

Dokuchits, E. V., Khasin, A. V., and Khassin, A. A. (2012). Mechanism and kinetics of hydrogen oxidation on silver. *Russian Chemical Bulletin*, 61(12):2225–2229.

Dovi, V. G., Reverberi, A. P., and Acevedo-Duarte, L. (1994). New procedure for optimal design of sequential experiments in kinetic models. *Industrial & Engineering Chemistry Research*, 33(1):62–68.

Echtermeyer, A., Amar, Y., Zakrzewski, J., and Lapkin, A. (2017). Self-optimisation and model-based design of experiments for developing a C-H activation flow process. *Beilstein Journal of Organic Chemistry*, 13:150–163.

Edgeworth, F. Y. (1887). On observations relating to several quantities. *Hermathena*, 6(13):279–285.

Edwards, K., Edgar, T. F., and Manousiouthakis, V. I. (2000). Reaction mechanism simplification using mixed-integer nonlinear programming. *Computers & Chemical Engineering*, 24(1):67–79.

Elliott, C., Vijayakumar, V., Zink, W., and Hansen, R. (2007). National Instruments LabVIEW: A Programming Environment for Laboratory Automation and Measurement. *JALA: Journal of the Association for Laboratory Automation*, 12(1):17–24.

Engl, H. W., Hanke, M., and Neubauer, A. (2000). *Regularization of Inverse Problems*. Springer Science & Business Media.

Engle, R. F. (1982). A general approach to lagrange multiplier model diagnostics. *Journal of Econometrics*, 20(1):83–104.

Engle, R. F. (1984). Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, volume 2, pages 775–826. Elsevier.

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

Espie, D. M. and Macchietto, S. (1988). Nonlinear transformations for parameter estimation. *Industrial & Engineering Chemistry Research*, 27(11):2175–2179.

F-Chart Software EES (2017). Engineering equation solver version 10.2. `http://fchartsoftware.com/ees/index.php/`.

Fabry, D. C., Sugiono, E., and Rueping, M. (2014). Self-Optimizing Reactor Systems: Algorithms, On-line Analytics, Setups, and Strategies for Accelerating Continuous Flow Process Optimization. *Israel Journal of Chemistry*, 54(4):341–350.

Fedorov, V. V. (1972). *Theory Of Optimal Experiments*. Academic Press, 1972 edition.

Fedorov, V. V. and Leonov, S. L. (2013). *Optimal Design for Nonlinear Response Models*. CRC Press, Boca Raton, 1 edition edition.

Fedorov, V. V. and Pázman, A. (1968). Design of Physical Experiments (Statistical Methods). *Fortschritte der Physik*, 16(6):325–355.

Fisher, R. A. (1992). Statistical Methods for Research Workers. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics: Methodology and Distribution*, Springer Series in Statistics, pages 66–70. Springer, New York, NY.

Florin Metenidis, M., Witczak, M., and Korbicz, J. (2004). A novel genetic programming approach to nonlinear system modelling: application to the DAMADICS benchmark problem. *Engineering Applications of Artificial Intelligence*, 17(4):363–370.

Floudas, C. A. and Pardalos, P. M. (2013). *Frontiers in Global Optimization*. Springer Science & Business Media.

Foerster, S., Unemo, M., Hathaway, L. J., Low, N., and Althaus, C. L. (2016). Time-kill curve analysis and pharmacodynamic modelling for in vitro evaluation of antimicrobials against Neisseria gonorrhoeae. *BMC Microbiology*, 16(1):216.

Fogler, H. S. (2005). *Elements of Chemical Reaction Engineering*. Prentice Hall, Upper Saddle River, NJ.

Fotopoulos, J., Georgakis, C., and Stenger, H. G. (1996). Effect of process-model mismatch on the optimization of the catalytic epoxidation of oleic acid using Tendency models. *Chemical Engineering Science*, 51(10):1899–1908.

Franceschini, G. and Macchietto, S. (2008a). Anti-Correlation Approach to Model-Based Experiment Design: Application to a Biodiesel Production Process. *Industrial & Engineering Chemistry Research*, 47(7):2331–2348.

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

Franceschini, G. and Macchietto, S. (2008c). Novel anticorrelation criteria for design of experiments: Algorithm and application. *AIChE Journal*, 54(12):3221–3238.

Franceschini, G. and Macchietto, S. (2008d). Novel anticorrelation criteria for model-based experiment design: Theory and formulations. *AIChE Journal*, 54(4):1009–1024.

Galvanin, F. (2010). *Optimal model-based design of experiments in dynamic systems: novel techniques and unconventional applications*. Ph.D. Thesis, University of Padova, Padova.

Galvanin, F., Barolo, M., and Bezzo, F. (2013). On the use of continuous glucose monitoring systems to design optimal clinical tests for the identification of type 1 diabetes models. *Computer Methods and Programs in Biomedicine*, 109(2):157–170.

Galvanin, F., Barolo, M., Padrini, R., Casonato, A., and Bezzo, F. (2014). A model-based approach to the automatic diagnosis of von Willebrand disease. *AIChE Journal*, 60(5):1718–1727.

Galvanin, F., Barolo, M., Pannocchia, G., and Bezzo, F. (2011). A disturbance estimation approach for online model-based redesign of experiments in the presence of systematic errors. *Computer Aided Chemical Engineering*, 29:467–471.

Galvanin, F., Barolo, M., Pannocchia, G., and Bezzo, F. (2012). Online model-based redesign of experiments with erratic models: A disturbance estimation approach. *Computers & Chemical Engineering*, 42:138–151.

Galvanin, F., Cao, E., Al-Rifai, N., Dua, V., and Gavriilidis, A. (2015). Optimal design of experiments for the identification of kinetic models of methanol oxidation over silver catalyst. *Chimica Oggi-Chemistry Today*, 33(3):51–56.

Galvanin, F., Macchietto, S., and Bezzo, F. (2007). Model-based design of parallel experiments. *Industrial & Engineering Chemistry Research*, 46(3).

Galvanin, F., Sankar, M., Cattaneo, S., Bethell, D., Dua, V., Hutchings, G. J., and Gavriilidis, A. (2018). On the development of kinetic models for solvent-free benzyl alcohol oxidation over a gold-palladium catalyst. *Chemical Engineering Journal*, 342:196–210.

Gandomi, A. H. and Alavi, A. H. (2011). Multi-stage genetic programming: A new strategy to nonlinear system modeling. *Information Sciences*, 181(23):5227–5239.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185(Supplement C):1–17.

Goodell, J. R., McMullen, J. P., Zaborenko, N., Maloney, J. R., Ho, C.-X., Jensen, K. F., Porco, J. A., and Beeler, A. B. (2009). Development of an Automated Microfluidic Reaction Platform for Multidimensional Screening: Reaction Discovery Employing Bicyclo[3.2.1]octanoid Scaffolds. *The Journal of Organic Chemistry*, 74(16):6169–6180.

Green, S. B., Thompson, M. S., and Poirier, J. (1999). Exploratory analyses to improve model fit: Errors due to misspecification and a strategy to reduce their occurrence. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):113–126.

Gromski, P. S., Henson, A. B., Granda, J. M., and Cronin, L. (2019). How to explore chemical space using algorithms and automation. *Nature Reviews Chemistry*, 3(2):119–128.

Hampel, F. R. (1985). The Breakdown Points of the Mean Combined With Some Rejection Rules. *Technometrics*, 27(2):95–107.

Hancock, G. and Compton, R. G., editors (1999). *Applications of Kinetic Modelling*. Elsevier Science, 1st edition.

Hansen, P. C. (2005). *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM.

Higham, N. J. (1996). *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Hill, P. D. H. (1978). A Review of Experimental Design Procedures for Regression Model Discrimination. *Technometrics*, 20(1):15–21.

Hill, R. W. (1977). *Robust regression when there are outliers in the carriers*. PhD thesis, Harvard University.

Hindmarsh, A. (1992). Odepack. a collection of ode system solvers. Technical report, Lawrence Livermore National Lab., CA, United States.

Hindmarsh, A. (2001). Odepack - a library of ode solvers written in fortran. `http://www.netlib.org/odepack/`.

Hof, P. M. J. v. d., Scherer, C., and Heuberger, P. S. C. (2009). *Model-Based Control - Bridging Rigorous Theory and Advanced Technology*. Springer Science & Business Media.

Holmes, N., Akien, G. R., Savage, R. J. D., Stanetty, C., Baxendale, I. R., Blacker, A. J., Taylor, B. A., Woodward, R. L., Meadows, R. E., and Bourne, R. A. (2016). Online quantitative mass spectrometry for the rapid adaptive optimisation of automated flow reactors. *Reaction Chemistry & Engineering*, 1(1):96–100.

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

Hsiang, T. and Reilly, P. M. (1971). A practical method for discriminating among mechanistic models. *The Canadian Journal of Chemical Engineering*, 49(6):865–871.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Huber, P. J. (2004). *Robust Statistics*. John Wiley & Sons, Inc., New York.

Hubert, M., Rousseeuw, P. J., and van Aelst, S. (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science*, 23(1):92–119.

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

Insel, P. A., Liljenquist, J. E., Tobin, J. D., Sherwin, R. S., Watkins, P., Andres, R., and Berman, M. (1975). Insulin Control of Glucose Metabolism in Man. *Journal of Clinical Investigation*, 55(5):1057–1066.

Jackson, J. E. (2003). *A User's Guide to Principal Components*. Wiley-Interscience, Hoboken, N.J.

Jeraal, M. I., Holmes, N., Akien, G. R., and Bourne, R. A. (2018). Enhanced process development using automated continuous reactors by self-optimisation algorithms and statistical empirical modelling. *Tetrahedron*, 74(25):3158–3164.

Johansen, T. A. (1997). On Tikhonov regularization, bias and variance in nonlinear system identification. *Automatica*, 33(3):441–446.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. `http://www.scipy.org/`.

Kahrs, O. and Marquardt, W. (2007). The validity domain of hybrid models and its application in process optimization. *Chemical Engineering and Processing: Process Intensification*, 46(11):1054–1066.

Karlsson, J., Anguelova, M., and Jirstrand, M. (2012). An Efficient Method for Structural Identifiability Analysis of Large Dynamic Systems. *IFAC Proceedings Volumes*, 45(16):941–946.

Katare, S., Caruthers, J. M., Delgass, W. N., and Venkatasubramanian, V. (2004). An Intelligent System for Reaction Kinetic Modeling and Catalyst Design. *Industrial & Engineering Chemistry Research*, 43(14):3484–3512.

King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9:137–163.

Klein, M. T., Hou, G., Bertolacini, R., Broadbelt, L. J., and Kumar, A. (2005). *Molecular Modeling in Heavy Hydrocarbon Conversions*. CRC Press, Boca Raton, 1st edition.

Klinke, D. J. I. and Wang, Q. (2017). Inferring the Impact of Regulatory Mechanisms that Underpin CD8+ T Cell Control of B16 Tumor Growth In vivo Using Mechanistic Models and Simulation. *Frontiers in Pharmacology*, 7.

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

Kristensen, N. R., Madsen, H., and Jørgensen, S. B. (2004). Stochastic Grey-Box Modelling as a Tool for Improving the Quality of First Engineering Principles Models. *IFAC Proceedings Volumes*, 37(1):143–148.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lee, P. L., Sullivan, G. R., and Zhou, W. (1989). Process/Model Mismatch Compensation for Model-Based Controllers. *Chemical Engineering Communications*, 80(1):33–51.

Levenspiel, O. (1998). *Chemical Reaction Engineering*. John Wiley & Sons, New York, 3rd edition.

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

López Cárdenas, D. C., Barz, T., Körkel, S., and Wozny, G. (2015). Nonlinear ill-posed problem analysis in model-based parameter estimation and experimental design. *Computers & Chemical Engineering*, 77:24–42.

Lu, Z. and Yang, W. (2004). Reaction path potential for complex systems derived from combined ab initio quantum mechanical and molecular mechanical calculations. *Journal of Chemical Physics*, 121:89–100.

MacKay, D. J. C. (1992). *Bayesian methods for adaptive models*. Ph.D. Thesis, California Institute of Technology.

Magoon, G. R. and Green, W. H. (2013). Design and implementation of a next-generation software interface for on-the-fly quantum and force field calculations in automated reaction mechanism generation. *Computers & Chemical Engineering*, 52:35–45.

Maheshwari, V., Rangaiah, G. P., and Samavedham, L. (2013). Multiobjective Framework for Model-based Design of Experiments to Improve Parameter Precision and Minimize Parameter Correlation. *Industrial & Engineering Chemistry Research*, 52(24):8289–8304.

Malig, T. C., Koenig, J. D. B., Situ, H., Chehal, N. K., Hultin, P. G., and Hein, J. E. (2017). Real-time HPLC-MS reaction progress monitoring using an automated analytical platform. *Reaction Chemistry & Engineering*, 2(3):309–314.

Mallows, C. (1975). On some topics in robustness: Technical memorandum. *Murray Hill, New Jersey: Bell Telephone Laboratories*.

Marin, G. B. and Yablonsky, G. S. (2011). *Kinetics of chemical reactions : decoding complexity*. Wiley-VCH, Weinheim.

Maronna, R., Bustos, O., and Yohai, V. (1979). Bias-and efficiency-robustness of general m-estimators for regression with random carriers. In *Smoothing techniques for curve estimation*, pages 91–116. Springer.

Marquardt, W. (2005). Model-Based Experimental Analysis of Kinetic Phenomena in Multi-Phase Reactive Systems. *Chemical Engineering Research and Design*, 83(6):561–573.

Mathworks MATLAB (2015). Matlab, version r2015a. `https://uk.mathworks.com/products/matlab.html`.

Mayorov, N., Gommers, R., Flamm, M., and Hagen, D. (2018). Lsoda - scipy wrapper to the fortran solver odepack. `https://github.com/scipy/scipy/blob/master/scipy/integrate/_ivp/lsoda.py`.

McMullen, J. P. and Jensen, K. F. (2010). An Automated Microfluidic System for Online Optimization in Chemical Synthesis. *Organic Process Research & Development*, 14(5):1169–1176.

McMullen, J. P. and Jensen, K. F. (2011). Rapid determination of reaction kinetics with an automated microfluidic system. *Organic Process Research & Development*, 15(2):398–407.

Meneghetti, N., Facco, P., Bezzo, F., and Barolo, M. (2014). A Methodology to Diagnose Process/Model Mismatch in First-Principles Models. *Industrial & Engineering Chemistry Research*, 53(36):14002–14013.

Mesbah, A. and Streif, S. (2015). A Probabilistic Approach to Robust Optimal Experiment Design with Chance Constraints. *IFAC-PapersOnLine*, 48(8):100–105.

Mizan, T. I. and Klein, M. T. (1999). Computer-assisted mechanistic modeling of n-hexadecane hydroisomerization over various bifunctional catalysts. *Catalysis Today*, 50(1):159–172.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Moore, J. S. and Jensen, K. F. (2012). Automated Multitrajectory Method for Reaction Optimization in a Microfluidic System using Online IR Analysis. *Organic Process Research & Development*, 16(8):1409–1415.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313.

Nelson, B. L. (1995). *Stochastic Modeling: Analysis & Simulation*. Courier Corporation.

Neumann, P., Cao, L., Russo, D., Vassiliadis, V. S., and Lapkin, A. A. (2019). A new formulation for symbolic regression to identify physico-chemical laws from experimental data. *Chemical Engineering Journal*, page 123412.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer, New York.

Nourbakhsh, M., Morris, N., Bergin, M., Iorio, F., and Grandi, D. (2016). Embedded sensors and feedback loops for iterative improvement in design synthesis for additive manufacturing. In *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers Digital Collection.

Ogunnaike, B. A. and Ray, W. H. (1994). *Process Dynamics, Modeling, and Control*. Oxford University Press, New York.

Oliphant, T. E. (2015). *Guide to NumPy*. CreateSpace Independent Publishing Platform, USA, 2nd edition.

Oliveira, L. P. d., Hudebine, D., Guillaume, D., and Verstraete, J. J. (2016). A Review of Kinetic Modeling Methodologies for Complex Processes. *Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles*, 71(3):45.

Özyurt, D. B. and Pike, R. W. (2004). Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Computers & Chemical Engineering*, 28(3):381–402.

Pardalos, P. M. and Resende, M. G. C., editors (2002). *Handbook of Applied Optimization*. Oxford University Press, New York, N.Y, 1st edition.

Parr, R. G. (1980). Density Functional Theory of Atoms and Molecules. In Fukui, K. and Pullman, B., editors, *Horizons of Quantum Chemistry*, Académie Internationale Des Sciences Moléculaires Quantiques / International Academy of Quantum Molecular Science, pages 5–15. Springer Netherlands.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petzold, L. (1983). Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations. *SIAM Journal on Scientific and Statistical Computing*, 4(1):136–148.

Petzold, L. and Zhu, W. (1999). Model reduction for chemical kinetics: An optimization approach. *AIChE Journal*, 45(4):869–886.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682.

Pipus, G., Plazl, I., and Koloini, T. (2000). Esterification of benzoic acid in microwave tubular flow reactor. *Chemical Engineering Journal*, 76(3):239–245.

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

Popper, K. R. (2002). *The Logic of Scientific Discovery*. Psychology Press, Hove, East Sussex, United Kingdom.

Prasad, V. and Vlachos, D. G. (2008). Multiscale model and informatics-based optimal design of experiments: application to the catalytic decomposition of ammonia on ruthenium. *Industrial & Engineering Chemistry Research*, 47(17):6555–6567.

Pritchard, D. J. and Bacon, D. W. (1978). Prospects for reducing correlations among parameter estimates in kinetic models. *Chemical Engineering Science*, 33(11):1539–1543.

PSE gPROMS (1997-2017). Process Systems Enterprise, gPROMS. `http://www.psenterprise.com/gproms`.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, New York.

Python Core Team (2018). Python: A dynamic, open source programming language. `https://www.python.org/`.

Pyzara, A., Bylina, B., and Bylina, J. (2011). The influence of a matrix condition number on iterative methods' convergence. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 459–464.

Quaglio, M., Bezzo, F., Gavriilidis, A., Cao, E., Al-Rifai, N., and Galvanin, F. (2019). Identification of kinetic models of methanol oxidation on silver in the presence of uncertain catalyst behavior. *AIChE Journal*, 65(10):e16707.

Rangarajan, S., Bhan, A., and Daoutidis, P. (2010). Rule-Based Generation of Thermochemical Routes to Biomass Conversion. *Industrial & Engineering Chemistry Research*, 49(21):10459–10470.

Ranzi, E., Cuoci, A., Faravelli, T., Frassoldati, A., Migliavacca, G., Pierucci, S., and Sommariva, S. (2008). Chemical Kinetics of Biomass Pyrolysis. *Energy & Fuels*, 22(6):4292–4300.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44(1):50–57.

Rasch, A. and Bücker, H. M. (2010). EFCOSS: an interactive environment facilitating optimal experimental design. *ACM Transactions on Mathematical Software*, 37(2):37.

Rasmuson, A., Andersson, B., Olsson, L., and Andersson, R. (2014). *Mathematical Modeling in Chemical Engineering*. Cambridge University Press, United Kingdom ; New York.

Raue, A., Karlsson, J., Saccomani, M. P., Jirstrand, M., and Timmer, J. (2014). Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics*, 30(10):1440–1448.

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

Rimensberger, T. and Rippin, D. W. T. (1986). "Sequential experimental design for precise parameter estimation. 1. Use of reparameterization". Comments. *Industrial & Engineering Chemistry Process Design and Development*, 25(4):1042–1044.

Rosenblueth, A. and Wiener, N. (1945). The Role of Models in Science. *Philosophy of Science*, 12(4):316–321.

Rossi, D., Gargiulo, L., Valitov, G., Gavriilidis, A., and Mazzei, L. (2017). Experimental characterization of axial dispersion in coiled flow inverters. *Chemical Engineering Research and Design*, 120:159–170.

Roth, P. M. (1966). *Design of Experiments for Discrimination Among Rival Models*. PhD thesis, Princeton University.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York.

Rousseeuw, P. J., Van Aelst, S., Van Driessen, K., and Gulló, J. A. (2004). Robust Multivariate Regression. *Technometrics*, 46(3):293–305.

Saccomani, M. P., Audoly, S., Bellu, G., D'angio, L., and Cobelli, C. (1997). Global Identifiability of Nonlinear Model Parameters. *IFAC Proceedings Volumes*, 30(11):233–238.

Saccomani, M. P., Audoly, S., and D'Angiò, L. (2003). Parameter identifiability of nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632.

Sahinidis, N. V. (1996). BARON: A general purpose global optimization software package. *Journal of Global Optimization*, 8(2):201–205.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. Springer, Tokyo : Dordrecht ; Boston : Hingham, MA.

Saltelli, A., Tarantola, S., and Campolongo, F. (2000). Sensitivity Analysis as an Ingredient of Modeling. *Statistical Science*, 15(4):377–395.

Sargent, R. (2005). Process systems engineering: A retrospective view with questions for the future. *Computers & Chemical Engineering*, 29(6):1237–1241.

Sargent, R. W. H. (1967). Integrated design and optimization of processes. *Chemical Engineering Progress*, 63(9):71–78.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts; London, England.

Schubert, H., Tegtmeyer, U., and Schlögl, R. (1994). On the mechanism of the selective oxidation of methanol over elemental silver. *Catalysis Letters*, 28(2-4):383–395.

Schwaab, M., Lemos, L. P., and Pinto, J. C. (2008a). Optimum reference temperature for reparameterization of the Arrhenius equation. Part 2: Problems involving multiple reparameterizations. *Chemical Engineering Science*, 63(11):2895–2906.

Schwaab, M., Luiz Monteiro, J., and Carlos Pinto, J. (2008b). Sequential experimental design for model discrimination: Taking into account the posterior covariance matrix

of differences between model predictions. *Chemical Engineering Science*, 63(9):2408–2419.

Schwaab, M. and Pinto, J. C. (2007). Optimum reference temperature for reparameterization of the Arrhenius equation. Part 1: Problems involving one kinetic constant. *Chemical Engineering Science*, 62(10):2750–2764.

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

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Sedoglavic, A. (2002). A Probabilistic Algorithm to Test Local Algebraic Observability in Polynomial Time. *Journal of Symbolic Computation*, 33(5):735–755.

Seigneur, C., Stephanopoulos, G., and Carr, R. W. (1982). Dynamic sensitivity analysis of chemical reaction systems: A variational method. *Chemical Engineering Science*, 37(6):845–853.

Seyedzadeh Khanshan, F. and West, R. H. (2016). Developing detailed kinetic models of syngas production from bio-oil gasification using Reaction Mechanism Generator (RMG). *Fuel*, 163:25–33.

Shahmohammadi, A. and McAuley, K. B. (2019). Sequential model-based A- and V-optimal design of experiments for building fundamental models of pharmaceutical production processes. *Computers & Chemical Engineering*, 129:106504.

Sidoli, F. R., Mantalaris, A., and Asprey, S. P. (2005). Toward Global Parametric Estimability of a Large-Scale Kinetic Single-Cell Model for Mammalian Cell Cultures. *Industrial & Engineering Chemistry Research*, 44(4):868–878.

Siegel, A. F. (1982). Robust Regression Using Repeated Medians. *Biometrika*, 69(1):242–244.

Silvey, S. D. (1959). The Lagrangian Multiplier Test. *The Annals of Mathematical Statistics*, 30(2):389–407.

Silvey, S. D. (1975). *Statistical Inference*. Chapman and Hall, London.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Söderström, T. and Stoica, P. (1989). *System Identification*. Prentice Hall, New York.

Song, J. (2004). *Building robust chemical reaction mechanisms : next generation of automatic model construction software*. Ph.D. Thesis, Massachusetts Institute of Technology.

Sorenson, H. W. (1980). *Parameter estimation: Principles and problems*. M. Dekker, New York, 1st edition.

Stamati, I., Logist, F., Akkermans, S., Noriega Fernández, E., and Van Impe, J. (2016). On the effect of sampling rate and experimental noise in the discrimination between microbial growth models in the suboptimal temperature range. *Computers & Chemical Engineering*, 85:84–93.

Steiner, S., Wolf, J., Glatzel, S., Andreou, A., Granda, J. M., Keenan, G., Hinkley, T., Aragon-Camarasa, G., Kitson, P. J., Angelone, D., and Cronin, L. (2019). Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*, 363(6423):eaav2211.

Stromberg, A. J. (1993). Computation of High Breakdown Nonlinear Regression Parameters. *Journal of the American Statistical Association*, 88(421):237–244.

Thybaut, J. W., Sun, J., Olivier, L., Van Veen, A. C., Mirodatos, C., and Marin, G. B. (2011). Catalyst design based on microkinetic models: Oxidative coupling of methane. *Catalysis Today*, 159(1):29–36.

Toffolo, G., Bergman, R. N., Finegood, D. T., Bowden, C. R., and Cobelli, C. (1980). Quantitative estimation of beta cell sensitivity to glucose in the intact organism: a minimal model of insulin kinetics in the dog. *Diabetes*, 29(12):979–990.

Transtrum, M. K., Machta, B., Brown, K., Daniels, B. C., Myers, C. R., and Sethna, J. P. (2015). Sloppiness and Emergent Theories in Physics, Biology, and Beyond. *arXiv:1501.07668*.

Transtrum, M. K., Machta, B. B., and Sethna, J. P. (2010). Why are nonlinear fits so challenging? *Physical Review Letters*, 104(6).

Tsay, C., Pattison, R. C., Baldea, M., Weinstein, B., Hodson, S. J., and Johnson, R. D. (2017). A superstructure-based design of experiments framework for simultaneous domain-restricted model identification and parameter estimation. *Computers & Chemical Engineering*, 107:408–426.

Ugi, I., Bauer, J., Bley, K., Dengler, A., Dietz, A., Fontain, E., Gruber, B., Herges, R., Knauer, M., Reitsam, K., and Stein, N. (1993). Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angewandte Chemie International Edition in English*, 32(2):201–227.

United Nations (2016). At UN, global leaders commit to act on antimicrobial resistance. `https://news.un.org/en/story/2016/09/539912-un-global-leaders-commit-act-antimicrobial-resistance`.

Van de Vijver, R., Vandewiele, N., Bhoorasingh, P. L., Slakman, B. L., Khanshan, F. S., Reyniers, M.-F., Marin, G., West, R., and Van Geem, K. (2015a). Automatic mechanism and kinetic model generation: a perspective on best practices, recent advances, and future challenges. *International Journal of Chemical Kinetics*, 47(4):199–231.

Van de Vijver, R., Vandewiele, N. M., Vandeputte, A. G., Van Geem, K. M., Reyniers, M.-F., Green, W. H., and Marin, G. B. (2015b). Rule-based ab initio kinetic model for alkyl sulfide pyrolysis. *Chemical Engineering Journal*, 278:385–393.

Vandewiele, N. M., Van Geem, K. M., Reyniers, M.-F., and Marin, G. B. (2012). Genesys: Kinetic model construction using chemo-informatics. *Chemical Engineering Journal*, 207-208:526–538.

Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., and Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers & Chemical Engineering*, 27(3):293–311.

Versyck, K. J. and Van Impe, J. F. (1997). On the unicity of optimal experimental design solutions for parameter estimation of microbial kinetics. In *1997 European Control Conference (ECC)*, pages 3509–3514.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482.

Waldron, C., Cao, E., Cattaneo, S., Brett, G. L., Miedziak, P. J., Wu, G., Sankar, M., Hutchings, G. J., and Gavriilidis, A. (2019a). Three step synthesis of benzylacetone and 4-(4-methoxyphenyl)butan-2-one in flow using micropacked bed reactors. *Chemical Engineering Journal*, 377:119976.

Waldron, C., Pankajakshan, A., Quaglio, M., Cao, E., Galvanin, F., and Gavriilidis, A. (2019b). An autonomous microreactor platform for the rapid identification of kinetic models. *Reaction Chemistry & Engineering*, 4(9):1623–1636.

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. E. (2011). *Probability and Statistics for Engineers and Scientists*. Pearson, London, 9th edition.

Walsh, B., Hyde, J. R., Licence, P., and Poliakoff, M. (2005). The automation of continuous reactions in supercritical CO2: the acid-catalysed etherification of short chain alcohols. *Green Chemistry*, 7(6):456–463.

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

Walter, E. and Pronzato, L. (1997). *Identification of Parametric Models from Experimental Data*. Springer-Verlag, London.

Web of Science (2019). Core collection database. `http://apps.webofknowledge.com/`.

White, A., Tolman, M., Thames, H. D., Withers, H. R., Mason, K. A., and Transtrum, M. K. (2016). The Limitations of Model-Based Experimental Design and Parameter Estimation in Sloppy Systems. *PLOS Computational Biology*, 12(12).

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Wilson, A. D., Schultz, J. A., and Murphey, T. D. (2015). Trajectory Optimization for Well-Conditioned Parameter Estimation. *IEEE Transactions on Automation Science and Engineering*, 12(1):28–36.

Wilson, Z. and Sahinidis, N. V. (2016). Simultaneous Reaction Identification and Parameter Estimation. *AIChE Annual Meeting*.

Wilson, Z. T. and Sahinidis, N. V. (2017). The ALAMO approach to machine learning. *Computers & Chemical Engineering*, 106:785–795.

Wolfram Research, Inc. (2019). Mathematica, Version 12.0. `https://www.wolfram.com/`.

Xiao-lei Yuan, Yan Bai, and Ling Dong (2008). Identification of linear time-invariant, nonlinear and time varying dynamic systems using genetic programming. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pages 56–61.

Xiong, Q. and Jutan, A. (2002). Grey-box modelling and control of chemical processes. *Chemical Engineering Science*, 57:1027–1039.

Yin, S., Ding, S. X., Xie, X., and Luo, H. (2014). A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. *IEEE Transactions on Industrial Electronics*, 61(11):6418–6428.

Yu, L. X. (2008). Pharmaceutical Quality by Design: Product and Process Development, Understanding, and Control. *Pharmaceutical Research*, 25(4):781–791.

Zeleznik, A. J. and Roth, J. (1978). Demonstration of the insulin receptor in vivo in rabbits and its possible role as a reservoir for the plasma hormone. *The Journal of Clinical Investigation*, 61(5):1363–1374.

Zhang, D., Chanona, E. A. D.-R., Vassiliadis, V. S., and Tamburic, B. (2015). Analysis of green algal growth via dynamic model simulation and process optimization. *Biotechnology and Bioengineering*, 112(10):2025–2039.

Zullo, L. C. (1991). *Computer aided design of experiments : An engineering approach*. PhD thesis, University of London.

Appendix A

Online RP - Simulated case: additional information

Additional details are presented in this appendix regarding the simulated experimental campaigns performed on the benzoic acid esterification system both in the absence and in the presence of online RP. Information related to the campaign performed without reparametrisation, i.e. the non-RP campaign, is reported in Table A.1. Information on the campaign conducted keeping the online reparametrisation *active*, i.e. the RP campaign, is given in Table A.2. In Table A.1 and Table A.2 the following information is given: 1) experimental conditions adopted to collect the samples, i.e. inlet concentration of benzoic acid $C_{\mathrm{BA}}^{\mathrm{IN}}$, flowrate F and temperature T; 2) sampled concentration of ethyl benzoate at the outlet $C_{\mathrm{EB}}^{\mathrm{OUT}}$; 3) computed parameter estimates $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$; 4) the pre-exponential factor and activation energy derived from the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ as $A = e^{\hat{\theta}_1}$ and $E_a = 10^4 \cdot \hat{\theta}_2$; 5) the sum of squared residuals $\chi^2_{\hat{Y}}$ and the reference value χ^2_{ref} computed from a χ^2 distribution with degree of freedom equal to the number of samples minus the number of parameters and 95% of significance.

Table A.1: Simulated case: experimental campaign with online RP option *inactive*. Experimental conditions, sampled concentrations, estimated kinetic parameters $\hat{\theta}$ (and related Arrhenius constants) and information regarding the goodness-of-fit are reported for the 9 samples collected in the campaign.

Simulated Case - Online RP Inactive

Sample	Experimental conditions φ			Sample	Estimates $\hat{\theta}$		Arrhenius constants[1]		Goodness-of-fit[2]	
number	C_{BA}^{IN} [mol L^{-1}]	F [µL min^{-1}]	T [K]	C_{EB}^{OUT} [mol L^{-1}]	$\hat{\theta}_1$	$\hat{\theta}_2$	A [s^{-1}]	E_a [J mol^{-1} K^{-1}]	χ^2_Y	χ^2_{ref}
1	1.50	20.00	413.0	0.368	-	-	-	-	-	-
2	1.00	10.00	393.0	0.188	-	-	-	-	-	-
3	1.25	15.00	403.0	0.272	12.15	6.56	$1.89{\cdot}10^5$	$6.56{\cdot}10^4$	0.5	3.8
4	1.55	7.50	385.0	0.212	14.83	7.47	$2.76{\cdot}10^6$	$7.47{\cdot}10^4$	3.3	6.0
5	1.55	7.50	412.5	0.871	15.99	7.85	$8.77{\cdot}10^6$	$7.85{\cdot}10^4$	6.6	7.8
6	1.55	7.50	389.0	0.320	15.06	7.53	$3.48{\cdot}10^6$	$7.53{\cdot}10^4$	9.2	9.5
7	1.55	7.50	413.0	0.860	14.90	7.47	$2.95{\cdot}10^6$	$7.47{\cdot}10^4$	9.5	11.1
8	1.55	7.50	393.5	0.383	14.84	7.45	$2.78{\cdot}10^6$	$7.45{\cdot}10^4$	9.5	12.6
9	1.55	7.50	413.0	0.831	14.94	7.49	$3.08{\cdot}10^6$	$7.49{\cdot}10^4$	9.8	14.1

[1] Pre-exponential factor and activation energy are computed from θ_1 and θ_2 as $A = e^{\theta_1}$ and $E_a = \theta_2 \cdot 10^4$
[2] A χ^2_{sample} larger than χ^2_{ref} is an index of inappropriate modelling assumptions

Table A.2: Simulated case: experimental campaign with online RP option *active*. Experimental conditions, sampled concentrations, estimated kinetic parameters $\hat{\theta}$ (and related Arrhenius constants) and information regarding the goodness-of-fit are reported for the 9 samples collected in the campaign.

Simulated Case - Online RP Active

Sample	Experimental conditions φ			Sample	Estimates $\hat{\theta}$		Arrhenius constants[1]		Goodness-of-fit[2]	
number	C_{BA}^{IN} [mol L^{-1}]	F [µL min^{-1}]	T [K]	C_{EB}^{OUT} [mol L^{-1}]	$\hat{\theta}_1$	$\hat{\theta}_2$	A [s^{-1}]	E_a [J mol^{-1} K^{-1}]	χ^2_Y	χ^2_{ref}
1	1.50	20.00	413.0	0.400	-	-	-	-	-	-
2	1.00	10.00	393.0	0.178	-	-	-	-	-	-
3	1.25	15.00	403.0	0.248	16.44	8.01	$1.38{\cdot}10^7$	$8.01{\cdot}10^4$	0.5	3.8
4	1.55	7.50	413.0	0.872	16.61	8.06	$1.64{\cdot}10^7$	$8.06{\cdot}10^4$	0.5	6.0
5	1.55	7.50	392.5	0.356	15.60	7.72	$5.96{\cdot}10^6$	$7.72{\cdot}10^4$	1.0	7.8
6	1.55	7.50	389.5	0.294	15.72	7.76	$6.71{\cdot}10^6$	$7.76{\cdot}10^4$	1.1	9.5
7	1.55	7.50	413.0	0.870	15.72	7.76	$6.72{\cdot}10^6$	$7.76{\cdot}10^4$	1.1	11.1
8	1.55	7.50	413.0	0.857	15.59	7.71	$5.87{\cdot}10^6$	$7.71{\cdot}10^4$	1.5	12.6
9	1.55	7.50	390.3	0.319	15.39	7.64	$4.83{\cdot}10^6$	$7.64{\cdot}10^4$	1.8	14.1

[1] Pre-exponential factor and activation energy are computed from θ_1 and θ_2 as $A = e^{\theta_1}$ and $E_a = \theta_2 \cdot 10^4$
[2] A χ^2_{sample} larger than χ^2_{ref} is an index of inappropriate modelling assumptions

Appendix B

Online RP - Additional simulated cases

A total number of 20 experimental campaigns were simulated to further validate the results presented in the manuscript. This was done primarily to demonstrate that the performance achieved by the algorithm both in the RP and in the non-RP campaigns is insensitive to the choice of the dataset (i.e. it is insensitive to the choice of the random seed used to generate the experimental data in-silico).

The results obtained in the simulated campaigns are reported in Table B.1. Campaigns 1-10 were performed applying the online reparametrisation method (RP campaigns), while campaigns 11-20 were performed without online reparametrisation (non-RP campaigns). As one can see from Table B.1, the algorithm with online RP option active retrieved the target parameter value in all the campaigns, i.e. the final p-value of the target parameters is above 1.00% in campaigns 1-10. The condition number of the log-likelihood functions at the end of experimental campaigns 1-10 is 1.0, demonstrating that the application of the online RP led to the elimination of the model sloppiness. In the campaigns where the online RP is inactive, i.e. campaigns 11-20, the final p-value is 0.00%, demonstrating the failure of the algorithm in retrieving the target value of the parameters. The failure is associated to the high condition number of the log-likelihood function, which is around $10^3 - 10^4$ in campaigns 11-20.

Table B.1: Results obtained in 20 simulated experimental campaigns: experimental campaigns 1-10 were performed keeping the online reparametrisation option *active*; campaigns 11-20 were performed keeping the option for online reparametrisation *inactive*. The p-value of the target parameters $\theta^* = [15.27, 7.6]$ given the final parameter statistics is reported together with the condition number of the log-likelihood function at the end of the experimental campaigns.

Campaign number	Online reparametrisation	Final p-value of target parameters θ^*	Final condition number κ
1	Active	64.74%	$1.0 \cdot 10^0$
2	Active	98.91%	$1.0 \cdot 10^0$
3	Active	91.98%	$1.0 \cdot 10^0$
4	Active	20.59%	$1.0 \cdot 10^0$
5	Active	30.52%	$1.0 \cdot 10^0$
6	Active	67.93%	$1.0 \cdot 10^0$
7	Active	16.61%	$1.0 \cdot 10^0$
8	Active	92.17%	$1.0 \cdot 10^0$
9	Active	23.19%	$1.0 \cdot 10^0$
10	Active	71.59%	$1.0 \cdot 10^0$
11	Inactive	0.00%	$9.6 \cdot 10^3$
12	Inactive	0.00%	$9.4 \cdot 10^3$
13	Inactive	0.00%	$9.5 \cdot 10^3$
14	Inactive	0.00%	$9.3 \cdot 10^3$
15	Inactive	0.00%	$1.1 \cdot 10^4$
16	Inactive	0.00%	$9.5 \cdot 10^3$
17	Inactive	0.00%	$1.0 \cdot 10^4$
18	Inactive	0.00%	$9.2 \cdot 10^3$
19	Inactive	0.00%	$8.7 \cdot 10^3$
20	Inactive	0.00%	$9.3 \cdot 10^3$

Appendix C

Online RP - Real case: additional information

Additional details are presented in this appendix regarding the non-RP campaign and the RP campaign performed on the experimental automated system. Information related to the campaign performed keeping the option for online model reparametrisation *inactive*, i.e. the non-RP campaign, is reported in Table C.1. Information on the campaign conducted keeping the option for online reparametrisation *active*, i.e. the RP campaign, is given in Table C.2. In Table C.1 and Table C.2 the following information is presented: 1) experimental conditions adopted to collect the samples, i.e. inlet concentration of benzoic acid $C_{\mathrm{BA}}^{\mathrm{IN}}$, flowrate F and temperature T; 2) sampled concentration of ethyl benzoate at the outlet $C_{\mathrm{EB}}^{\mathrm{OUT}}$; 3) parameter estimates $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2]$ returned by the model identification algorithm; 4) the pre-exponential factor and activation energy computed from the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ as $A = e^{\hat{\theta}_1}$ and $E_a = 10^4 \cdot \hat{\theta}_2$; 5) the sum of squared residuals χ^2_Y and the reference value χ^2_{ref} computed from a χ^2 distribution with degree of freedom equal to the number of samples minus the number of parameters and 95% of significance.

A sum of squared residuals χ^2_Y larger than the reference value χ^2_{ref} is interpreted as an index of inappropriate modelling assumptions (Silvey, 1975). As one can see from Table C.1, the χ^2_Y after the collection of 9 samples in the non-RP campaign is 5.92. From Table C.2, it can be appreciated that the χ^2_Y after the collection of 9 samples in the RP campaign is 1.83. Both in the non-RP and in the RP campaign the χ^2_Y is smaller than the $\chi^2_{ref} = 17.88$, thus demonstrating that the modelling assumptions (see Section 3.3.2) are not falsified by the experimental evidence.

As one can see from Table C.1, the experimental conditions designed by the algorithm

for samples 5, 7 and 9 in the non-RP case were similar, i.e. inlet concentration of benzoic acid $C_{BA}^{IN} = 1.55\ \text{mol}\,\text{L}^{-1}$, flowrate around $F = 7.5\ \mu\text{L}\,\text{min}^{-1}$ and temperature $T = 413.0$ K, i.e. the upper limit for the temperature. Samples 4, 6 and 8 were instead designed by the algorithm at conditions $C_{BA}^{IN} = 1.55\ \text{mol}\,\text{L}^{-1}$, flowrate $F = 7.5\ \mu\text{L}\,\text{min}^{-1}$ and temperature in the range $T = 383.0 - 390.0$ K. The designed samples in the non-RP case suggest the presence of two optimally informative sets of experimental conditions at maximum temperature $T = 413.0$ K and at temperature around $T = 385.0$ K, given that the inlet concentration of benzoic acid C_{BA}^{IN} is set at the maximum and that flowrate F is set at the minimum.

An analogous situation can be observed in the RP case. As one can see from Table C.2, samples 4, 7 and 9 in the RP case were designed at conditions $C_{BA}^{IN} = 1.55\ \text{mol}\,\text{L}^{-1}$, $F = 7.5\ \mu\text{L}\,\text{min}^{-1}$ and $T = 413.0$ K. Samples 5, 6 and 8 were instead designed at conditions $C_{BA}^{IN} = 1.55\ \text{mol}\,\text{L}^{-1}$, $F = 7.5\ \mu\text{L}\,\text{min}^{-1}$ and temperature around $T = 391.0$ K.

Table C.1: Real case: experimental campaign with online RP option *inactive*. Experimental conditions, sampled concentrations, estimated kinetic parameters $\hat{\theta}$ (and related Arrhenius constants) and information regarding the goodness-of-fit are reported for the 9 samples collected in the campaign.

Real Case - Online RP Inactive

Sample	Experimental conditions φ			Sample	Estimates $\hat{\theta}$		Arrhenius constants[1]		Goodness-of-fit[2]	
number	C_{BA}^{IN} [mol L^{-1}]	F [μL min^{-1}]	T [K]	C_{EB}^{OUT} [mol L^{-1}]	$\hat{\theta}_1$	$\hat{\theta}_2$	A [s^{-1}]	E_a [J mol^{-1} K^{-1}]	χ^2_Y	χ^2_{ref}
1	1.50	20.00	413.0	0.370	-	-	-	-	-	-
2	1.00	10.00	393.0	0.161	-	-	-	-	-	-
3	1.25	15.00	403.0	0.240	16.16	7.94	$1.04{\cdot}10^{7}$	$7.94{\cdot}10^{4}$	$4.35{\cdot}10^{-4}$	3.84
4	1.55	7.50	383.0	0.175	16.44	8.03	$1.39{\cdot}10^{7}$	$8.03{\cdot}10^{4}$	$2.65{\cdot}10^{-2}$	5.99
5	1.55	7.58	413.0	0.848	17.15	8.26	$2.81{\cdot}10^{7}$	$8.26{\cdot}10^{4}$	1.04	7.81
6	1.55	7.50	390.2	0.284	16.80	8.14	$1.98{\cdot}10^{7}$	$8.14{\cdot}10^{4}$	1.31	9.49
7	1.55	7.50	413.0	0.876	17.23	8.28	$3.03{\cdot}10^{7}$	$8.28{\cdot}10^{4}$	3.56	11.07
8	1.55	7.50	388.5	0.254	17.15	8.26	$2.82{\cdot}10^{7}$	$8.26{\cdot}10^{4}$	3.59	12.59
9	1.55	7.50	413.0	0.887	17.42	8.34	$3.69{\cdot}10^{7}$	$8.34{\cdot}10^{4}$	5.92	14.07

[1] Pre-exponential factor and activation energy are computed from θ_1 and θ_2 as $A = e^{\theta_1}$ and $E_a = 10^4 \cdot \theta_2$
[2] A χ^2_Y larger than χ^2_{ref} is an index of inappropriate modelling assumptions

Table C.2: Real case: experimental campaign with online RP option *active*. Experimental conditions, sampled concentrations, estimated kinetic parameters $\hat{\theta}$ (and related Arrhenius constants) and information regarding the goodness-of-fit are reported for the 9 samples collected in the campaign.

Real Case - Online RP Active

Sample	Experimental conditions φ			Sample	Estimates $\hat{\theta}$		Arrhenius constants[1]		Goodness-of-fit[2]	
number	C_{BA}^{IN} [mol L^{-1}]	F [µL min^{-1}]	T [K]	C_{EB}^{OUT} [mol L^{-1}]	$\hat{\theta}_1$	$\hat{\theta}_2$	A [s^{-1}]	E_a [J mol^{-1} K^{-1}]	χ^2_Y	χ^2_{ref}
1	1.50	20.00	413.0	0.409	-	-	-	-	-	-
2	1.00	10.00	393.0	0.172	-	-	-	-	-	-
3	1.25	15.00	403.0	0.252	17.54	8.37	$4.13{\cdot}10^7$	$8.37{\cdot}10^4$	0.21	3.84
4	1.55	7.50	413.0	0.900	18.12	8.56	$7.39{\cdot}10^7$	$8.56{\cdot}10^4$	0.52	5.99
5	1.55	7.50	392.3	0.346	16.86	8.13	$2.10{\cdot}10^7$	$8.13{\cdot}10^4$	1.27	7.81
6	1.55	7.50	390.6	0.307	16.90	8.15	$2.18{\cdot}10^7$	$8.15{\cdot}10^4$	1.27	9.49
7	1.55	7.50	413.0	0.895	16.91	8.15	$2.20{\cdot}10^7$	$8.15{\cdot}10^4$	1.27	11.07
8	1.55	7.50	391.2	0.323	16.83	8.12	$2.04{\cdot}10^7$	$8.12{\cdot}10^4$	1.31	12.59
9	1.55	7.50	413.0	0.908	16.98	8.17	$2.36{\cdot}10^7$	$8.17{\cdot}10^4$	1.83	14.07

[1] Pre-exponential factor and activation energy are computed from θ_1 and θ_2 as $A = e^{\theta_1}$ and $E_a = 10^4 \cdot \theta_2$

[2] A χ^2_Y larger than χ^2_{ref} is an index of inappropriate modelling assumptions

Appendix D

Ethanol dehydrogenation - Experimental data generated in-silico

Additional information is provided regarding the simulated experimental campaign conducted on the ethanol dehydrogenation system illustrated in Section 4.3.1. The experimental campaign is simulated in an ideal tubular reactor of unit length assuming a fixed catalyst mass $w = 2.0$ g. The LHHW kinetic model proposed by Carotenuto et al. (2013) is employed for the generation of the in-silico data. For the generation of the dataset, the kinetic parameters are set to the values reported in (Carotenuto et al., 2013). Numerical values and associated units are also reported in Table D.1. The values for the equilibrium constants are computed from the following Van't Hoff equations

$$K_{eq1} = e^{16.5-9134.6/T} \tag{D.1}$$

$$K_{eq2} = e^{-4.79+4386.0/T} \tag{D.2}$$

The experimental conditions observed in the course of the experimental campaign are reported in Table D.2 together with the associated sampled values at the outlet of the tubular reactor. Samples 1-8 were obtained with a full factorial design with three factors (i.e. ethanol inlet flowrate, total pressure and temperature) and two levels for each factor. These were the samples analysed in the ML case.

The additional samples, i.e. samples 9-16, were collected in the MBDM case adopting an A-optimal MBDoE criterion constrained within the model reliability domain. The value for the binary switchers β computed at the last iteration of the experimental campaign in

the MBDM case is also reported in Table D.2 for all the generated samples.

Table D.1: Parameter values of the LHHW model of ethanol dehydrogenation on copper-based catalyst (from Carotenuto et al. (2013)).

Parameter	Value	Unit
A_1	$1.13 \cdot 10^{18}$	[mol g^{-1} h^{-1}]
A_2	$4.87 \cdot 10^{4}$	[mol g^{-1} h^{-1}]
A_3	$1.0 \cdot 10^{-3}$	[mol g^{-1} h^{-1} bar^{-2}]
E_{a1}	$1.52 \cdot 10^{5}$	[J mol^{-1} K^{-1}]
E_{a2}	$5.42 \cdot 10^{4}$	[J mol^{-1} K^{-1}]
E_{a3}	$6.70 \cdot 10^{-1}$	[J mol^{-1} K^{-1}]
b_{EtOH}	10.40	[bar^{-1}]
b_{AcH}	98.40	[bar^{-1}]
b_{EA}	41.20	[bar^{-1}]
$b_{\mathrm{H_2}}$	$2.50 \cdot 10^{-4}$	[bar^{-1}]

Table D.2: Simulated experimental campaign conducted on the ethanol dehydrogenation system in a tubular reactor with unit length assuming a fixed catalyst mass $w = 2.0$ g.

Sample	Inlet molar flowrate [mol h^{-1}]					Press. [bar]	Temp. [K]	Measured Outlet Flowrates [mol h^{-1}]				Computed
number	EtOH	AcH	EA	H_2	N_2	P_{TOT}	T	EtOH	AcH	EA	H_2	$\hat{\beta}$
1	0.100	0.0	0.0	0.057	0.057	10.0	453.15	0.126	0.033	0.031	0.066	+1
2	2.500	0.0	0.0	0.057	0.057	10.0	453.15	2.456	0.036	0.022	0.135	-1
3	0.100	0.0	0.0	0.057	0.057	30.0	453.15	0.111	0.013	0.009	0.052	+1
4	2.500	0.0	0.0	0.057	0.057	30.0	453.15	2.490	0.022	0.011	0.054	-1
5	0.100	0.0	0.0	0.057	0.057	10.0	533.15	0.037	0.010	0.017	0.108	+1
6	2.500	0.0	0.0	0.057	0.057	10.0	533.15	1.839	0.450	0.093	0.718	+1
7	0.100	0.0	0.0	0.057	0.057	30.0	533.15	0.059	0.003	0.019	0.112	+1
8	2.500	0.0	0.0	0.057	0.057	30.0	533.15	2.071	0.206	0.111	0.510	+1
9	1.242	0.0	0.0	0.057	0.057	18.9	512.29	1.027	0.062	0.061	0.262	+1
10	2.498	0.0	0.0	0.057	0.057	29.9	532.91	2.041	0.197	0.119	0.496	+1
11	0.578	0.0	0.0	0.057	0.057	18.9	515.18	0.444	0.039	0.051	0.188	+1
12	2.500	0.0	0.0	0.057	0.057	30.0	499.13	2.288	0.158	0.027	0.264	-1
13	1.674	0.0	0.0	0.057	0.057	25.5	506.24	1.464	0.097	0.042	0.246	-1
14	0.887	0.0	0.0	0.057	0.057	30.0	479.41	0.839	0.006	0.003	0.105	-1
15	1.767	0.0	0.0	0.057	0.057	30.0	517.00	1.515	0.089	0.079	0.314	+1
16	0.649	0.0	0.0	0.057	0.057	19.9	501.99	0.552	0.050	0.052	0.156	+1

Appendix E

Methanol oxidation - Experimental data

Table E.1 reports the experimental data collected in a campaign aimed at the identification of a kinetic model of methanol oxidation on silver catalyst. Data include: temperature T; inlet and outlet pressure P; inlet and outlet flowrates F (referred to standard conditions STC at temperature $T = 273.15$ K and pressure $P = 101325$ Pa); inlet and outlet molar fractions for methanol y_{CH_3OH}, oxygen y_{O_2}, water y_{H_2O}, formaldehyde y_{CH_2O}, hydrogen y_{H_2} and carbon dioxide y_{CO_2}.

Table E.1: Experimental data associated with the experimental campaign conducted on the microreactor platform illustrated in Section 4.3.2.

Sample	T [K]	Location	P [Pa]	F^* [ml min^{-1}]	y_{CH_3OH}	y_{O_2}	y_{H_2O}	y_{CH_2O}	y_{H_2}	y_{CO_2}
1	783	Inlet	260000	73.1	0.0994	0.0415	0.0753	0.0	0.0	0.0
		Outlet	160000	76.8	0.0124	0.0	0.1401	0.0749	0.0179	0.0078
2	783	Inlet	220000	41.7	0.0997	0.0414	0.0755	0.0	0.0	0.0
		Outlet	160000	44.0	0.0094	0.0	0.1468	0.075	0.0188	0.0071
3	783	Inlet	200000	29.1	0.0996	0.0414	0.0755	0.0	0.0	0.0
		Outlet	160000	30.3	0.0101	0.0	0.136	0.0748	0.0187	0.0065
4	733	Inlet	220000	50.9	0.1468	0.0975	0.2293	0.0	0.0	0.0
		Outlet	160000	53.5	0.0339	0.0101	0.3568	0.0447	0.0133	0.0391
5	765	Inlet	226000	50.9	0.1468	0.0975	0.2293	0.0	0.0	0.0
		Outlet	160000	53.6	0.0123	0.0006	0.3401	0.0893	0.0201	0.0359
6	796	Inlet	235000	50.9	0.1468	0.0975	0.2293	0.0	0.0	0.0
		Outlet	160000	53.7	0.0049	0.0002	0.3467	0.0998	0.0188	0.0293
7	826	Inlet	240000	50.9	0.1468	0.0975	0.2293	0.0	0.0	0.0
		Outlet	160000	53.8	0.0016	0.0001	0.3417	0.107	0.0195	0.0309
8	765	Inlet	280000	93.9	0.1469	0.098	0.2296	0.0	0.0	0.0
		Outlet	160000	99.8	0.0171	0.0063	0.3467	0.0865	0.0174	0.0306
9	796	Inlet	286000	93.9	0.1469	0.098	0.2296	0.0	0.0	0.0
		Outlet	160000	100.0	0.0054	0.0026	0.3481	0.1	0.0181	0.0312
10	826	Inlet	295000	93.9	0.1469	0.098	0.2296	0.0	0.0	0.0
		Outlet	160000	100.5	0.0015	0.0016	0.3507	0.1079	0.0179	0.0282
11	800	Inlet	240000	54.6	0.259	0.1064	0.2122	0.0	0.0	0.0
		Outlet	160000	59.7	0.043	0.0001	0.3449	0.1686	0.0375	0.0187
12	850	Inlet	245000	54.6	0.259	0.1064	0.2122	0.0	0.0	0.0
		Outlet	160000	59.8	0.026	0.0	0.3183	0.2089	0.0458	0.0143
13	900	Inlet	252000	54.6	0.259	0.1064	0.2122	0.0	0.0	0.0
		Outlet	160000	60.6	0.0187	0.0	0.3226	0.2116	0.0523	0.0127

* at temperature $T = 273.15$ K; pressure $P = 101325$ Pa.

Appendix F

Baker's yeast system - Experimental data generated in-silico

The in-silico dataset presented in Table F.1 was generated by integrating the Contois-type baker's yeast growth model presented in Section 5.3.1.1 using the parameter values $\theta^* = [0.310, 0.180, 0.550, 0.050]^T$. The 28 samples of biomass concentration x_1 [g L^{-1}] and substrate concentration x_2 [g L^{-1}] were generated adding uncorrelated Gaussian measurement noise with covariance $\Sigma_y = 2.5 \cdot 10^{-3}\mathbf{I}$. A full factorial design was adopted with 2 levels for the dilution factor $u_1 = \{0.05, 0.20\}$ [h^{-1}], 2 levels for the substrate concentration in the feed $u_2 = \{5.0, 35.0\}$ [g L^{-1}] and 7 levels for the sampling time $t = \{3.0, 6.0, 9.0, 12.0, 15.0, 18.0, 21.0\}$ [h].

The dataset in Table F.1 was employed in both in Section 5.3.1 and in Section 6.3.1 for the identification, diagnosis and evolution of the approximated Monod-type baker's yeast growth model.

Table F.1: Experimental conditions and samples generated in-silico for the simulated experimental campaign conducted on the baker's yeast system illustrated in Section 5.3.1.

Sample	Initial states $\mathbf{x}(0)$		Inputs $\mathbf{u}$		Sampling	Measured states $\mathbf{x}(t)$	
number	x_1 [g L^{-1}]	x_2 [g L^{-1}]	u_1 [h^{-1}]	u_2 [g L^{-1}]	time t [h]	x_1 [g L^{-1}]	x_2 [g L^{-1}]
1	1.00	0.01	0.05	5.00	3.00	1.12	0.15
2	1.00	0.01	0.05	5.00	6.00	1.17	0.24
3	1.00	0.01	0.05	5.00	9.00	1.27	0.08
4	1.00	0.01	0.05	5.00	12.00	1.27	0.11
5	1.00	0.01	0.05	5.00	15.00	1.25	0.14
6	1.00	0.01	0.05	5.00	18.00	1.28	0.19
7	1.00	0.01	0.05	5.00	21.00	1.33	0.12
8	1.00	0.01	0.05	35.00	3.00	1.64	3.20
9	1.00	0.01	0.05	35.00	6.00	2.88	4.49
10	1.00	0.01	0.05	35.00	9.00	4.66	3.71
11	1.00	0.01	0.05	35.00	12.00	6.58	1.75
12	1.00	0.01	0.05	35.00	15.00	7.77	0.91
13	1.00	0.01	0.05	35.00	18.00	8.31	0.80
14	1.00	0.01	0.05	35.00	21.00	8.51	0.84
15	1.00	0.01	0.20	5.00	3.00	1.04	1.38
16	1.00	0.01	0.20	5.00	6.00	1.06	1.83
17	1.00	0.01	0.20	5.00	9.00	1.11	1.88
18	1.00	0.01	0.20	5.00	12.00	1.24	1.96
19	1.00	0.01	0.20	5.00	15.00	1.42	1.90
20	1.00	0.01	0.20	5.00	18.00	1.43	1.68
21	1.00	0.01	0.20	5.00	21.00	1.47	1.49
22	1.00	0.01	0.20	35.00	3.00	1.06	14.59
23	1.00	0.01	0.20	35.00	6.00	1.33	22.14
24	1.00	0.01	0.20	35.00	9.00	1.55	26.12
25	1.00	0.01	0.20	35.00	12.00	1.83	27.87
26	1.00	0.01	0.20	35.00	15.00	2.21	28.48
27	1.00	0.01	0.20	35.00	18.00	2.64	28.26
28	1.00	0.01	0.20	35.00	21.00	3.13	27.72

Appendix G

Glucose-insulin interaction system - Experimental data generated in-silico

The simulated IVGTTs presented in the following Tables were obtained by integrating the glucose-insulin interaction model presented in Section 5.3.2.1 using the parameter values $\theta^* = [2.96 \cdot 10^{-2}, 6.51 \cdot 10^{-6}, 1.86 \cdot 10^{-2}, 5.36 \cdot 10^{-3}, 9.09 \cdot 10^{1}, 2.3 \cdot 10^{-1}]^T$. The in-silico data were generated adding uncorrelated Gaussian measurement noise as specified in the captions of each table. The data provided in this Appendix were used to identify and diagnose the approximated model of glucose-insulin interaction presented in Section 5.3.2.2. The IVGTT reported in Table G.1 (only measurements of G and I) was also analysed in Section 6.3.2 for computing the ERIs and support the evolution of two approximated models of glucose-insulin interaction.

Table G.1: IVGTT simulated on a healthy test subject with basal glucose concentration $G_b = 93.0$ [mg dL^{-1}]. The experimental design is the same proposed by Bergman et al. (1981) and involves 23 samples. The sampled quantities are G [mg dL^{-1}], $I(t)$ [μU mL^{-1}] and X [μU min mL^{-1}]. Gaussian noise was added with standard deviations 1.0 mg dL^{-1} for measurements of G, 1.5 μU mL^{-1} for measurements of I and 10.0 μU min mL^{-1} for X.

IVGTT 1							
Sample	Initial states			Sampling	Measured states		
number	$G(0)$	$X(0)$	$I(0)$	time t [h]	$G(t)$	$I(t)$	$X(t)$
1	298.00	0.00	333.00	2.00	286.93	213.51	528.55
2	298.00	0.00	333.00	4.00	273.90	137.32	861.95
3	298.00	0.00	333.00	6.00	259.17	95.00	1036.59
4	298.00	0.00	333.00	8.00	245.62	71.98	1161.12
5	298.00	0.00	333.00	10.00	233.59	56.22	1240.13
6	298.00	0.00	333.00	12.00	221.49	47.65	1309.36
7	298.00	0.00	333.00	14.00	210.51	40.39	1328.62
8	298.00	0.00	333.00	16.00	200.60	41.52	1381.99
9	298.00	0.00	333.00	19.00	188.10	42.16	1407.56
10	298.00	0.00	333.00	22.00	172.91	44.34	1483.14
11	298.00	0.00	333.00	27.00	154.80	40.12	1541.41
12	298.00	0.00	333.00	32.00	137.23	39.20	1590.24
13	298.00	0.00	333.00	42.00	117.05	28.86	1649.78
14	298.00	0.00	333.00	52.00	99.94	14.80	1560.09
15	298.00	0.00	333.00	62.00	88.73	3.72	1418.43
16	298.00	0.00	333.00	72.00	84.35	1.61	1164.47
17	298.00	0.00	333.00	82.00	80.26	1.29	976.66
18	298.00	0.00	333.00	92.00	81.04	1.76	807.71
19	298.00	0.00	333.00	102.00	80.44	0.10	679.16
20	298.00	0.00	333.00	122.00	81.97	0.54	458.88
21	298.00	0.00	333.00	142.00	82.95	1.21	317.11
22	298.00	0.00	333.00	162.00	83.90	0.60	222.85
23	298.00	0.00	333.00	182.00	85.75	1.36	157.03

Table G.2: IVGTT simulated on a healthy test subject with basal glucose concentration $G_b = 93.0$ [mg dL^{-1}]. The experimental design is the same proposed by Bergman et al. (1981) and involves 23 samples. The sampled quantities are G [mg dL^{-1}], $I(t)$ [μU mL^{-1}]. Gaussian noise was added with standard deviations 1.0 mg dL^{-1} for measurements of G and 1.5 μU mL^{-1} for measurements of I.

IVGTT 2						
Sample	Initial states			Sampling	Measured states	
number	$G(0)$	$X(0)$	$I(0)$	time t [h]	$G(t)$	$I(t)$
1	276.00	0.00	69.00	2.00	265.33	45.39
2	276.00	0.00	69.00	4.00	256.02	33.58
3	276.00	0.00	69.00	6.00	244.24	26.85
4	276.00	0.00	69.00	8.00	235.10	24.79
5	276.00	0.00	69.00	10.00	227.26	25.32
6	276.00	0.00	69.00	12.00	219.22	31.64
7	276.00	0.00	69.00	14.00	209.33	34.18
8	276.00	0.00	69.00	16.00	201.27	35.69
9	276.00	0.00	69.00	19.00	191.57	40.23
10	276.00	0.00	69.00	22.00	180.25	42.34
11	276.00	0.00	69.00	27.00	164.92	42.98
12	276.00	0.00	69.00	32.00	149.30	43.52
13	276.00	0.00	69.00	42.00	123.49	38.56
14	276.00	0.00	69.00	52.00	112.37	31.76
15	276.00	0.00	69.00	62.00	99.34	19.24
16	276.00	0.00	69.00	72.00	90.33	5.80
17	276.00	0.00	69.00	82.00	85.83	1.82
18	276.00	0.00	69.00	92.00	83.95	0.49
19	276.00	0.00	69.00	102.00	83.89	0.22
20	276.00	0.00	69.00	122.00	82.34	1.00
21	276.00	0.00	69.00	142.00	84.65	0.80
22	276.00	0.00	69.00	162.00	85.30	0.95
23	276.00	0.00	69.00	182.00	88.25	0.31

Table G.3: IVGTT simulated on a healthy test subject with basal glucose concentration $G_b = 93.0$ [mg dL^{-1}]. The experimental design is the same proposed by Bergman et al. (1981) and involves 23 samples. The sampled quantities are G [mg dL^{-1}], $I(t)$ [μU mL^{-1}]. Gaussian noise was added with standard deviations 5.0 mg dL^{-1} for measurements of G and 7.5 μU mL^{-1} for measurements of I.

IVGTT 3						
Sample	Initial states			Sampling	Measured states	
number	$G(0)$	$X(0)$	$I(0)$	time t [h]	$G(t)$	$I(t)$
1	298.00	0.00	333.00	2.00	293.98	219.39
2	298.00	0.00	333.00	4.00	282.87	131.45
3	298.00	0.00	333.00	6.00	262.97	94.39
4	298.00	0.00	333.00	8.00	247.26	80.71
5	298.00	0.00	333.00	10.00	236.63	58.88
6	298.00	0.00	333.00	12.00	222.82	46.42
7	298.00	0.00	333.00	14.00	211.76	25.08
8	298.00	0.00	333.00	16.00	203.22	37.07
9	298.00	0.00	333.00	19.00	197.18	42.43
10	298.00	0.00	333.00	22.00	172.16	53.16
11	298.00	0.00	333.00	27.00	155.42	34.79
12	298.00	0.00	333.00	32.00	129.31	40.14
13	298.00	0.00	333.00	42.00	121.97	26.54
14	298.00	0.00	333.00	52.00	98.73	6.28
15	298.00	0.00	333.00	62.00	81.90	0.67
16	298.00	0.00	333.00	72.00	82.59	6.28
17	298.00	0.00	333.00	82.00	73.81	6.67
18	298.00	0.00	333.00	92.00	82.59	8.85
19	298.00	0.00	333.00	102.00	80.33	0.49
20	298.00	0.00	333.00	122.00	83.18	2.72
21	298.00	0.00	333.00	142.00	80.27	6.09
22	298.00	0.00	333.00	162.00	76.99	3.01
23	298.00	0.00	333.00	182.00	79.22	6.80

Appendix H

Multivariate MMI-based analysis

In Section 5.2.3, a statistical test based on the Lagrange multipliers test was formulated with the aim of detecting which parameters should be considered for evolution in under-fitting parametric models. The test formulated in Section 5.2.3 did not consider the possible interaction between the parameter under diagnosis and the other model parameters. In this Appendix, a multivariate version of the same test is formulated and it is tested on the same case study presented in Section 5.3.1.

H.1 Lagrange multiplier test

A multivariate statistical test for diagnosing process-model mismatch is introduced with the aim of testing the hypothesis that a certain parameter θ_i is a state-independent constant. Without loss of generality, it is assumed that the parameter under analysis is $\theta_i = \theta_1$. The competing hypotheses under test are:

Null hypothesis H_0**.** θ_1 and $\theta_j \; \forall \, j \neq 1$ are all state-independent constants.

Alternative hypothesis H_a**.** θ_1 is a state-dependent function and $\theta_j \; \forall \, j \neq 1$ are state-independent constants.

The log-likelihood function is written assuming that θ_1 is a function of the experimental conditions φ, i.e., $\theta_1 = g(\varphi)$ (knowledge of the functional form of g is not required in the test). The parameter array Θ_m is defined as the $(N + N_\theta - 1) \times 1$ array of parameters $\Theta_m = [\theta_{1,1}, ..., \theta_{1,N}, \theta_2, \theta_3, ..., \theta_{N_\theta}]^T$. In Θ_m, parameter $\theta_{1,i} = g(\varphi_i)$ represents the value of function g at experimental conditions φ_i. Let $\mathscr{L}_m(Y|\Theta_m)$ be the log-likelihood function

written for dataset Y under parametrisation Θ_m.

$$\begin{aligned}\mathscr{L}_m(Y|\Theta_m) &= -\frac{N}{2}[N_y \ln(2\pi) + \ln(\det(\Sigma_y))] \\ &- \frac{1}{2}\sum_{i=1}^{N}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta_{1,i}, \theta_2, ..., \theta_{N_\theta})]^T \Sigma_y^{-1}[\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta_{1,i}, \theta_2, ..., \theta_{N_\theta})]\end{aligned} \tag{H.1}$$

Under parametrisation Θ_m, the i-th element in the sum in (H.1) is a function of parameters $\theta_{1,i}, \theta_2, ..., \theta_{N_\theta}$. One shall notice that in (5.1), the i-th element in the sum is a function of parameter $\theta_{1,i}$ only.

The set of $N-1$ functions $\mathbf{s}$ is defined as

$$\mathbf{s} = [\theta_{1,1} - \theta_{1,2}, ..., \theta_{1,N-1} - \theta_{1,N}]^T \tag{5.2}$$

As in Section 5.2.3, the null and alternative hypotheses are formalised as the presence/absence of an $N-1$ set of constraints for the functions $\mathbf{s}$ as follows

$$\begin{aligned} H_0: &\quad \mathbf{s} = \mathbf{0} \\ H_a: &\quad \mathbf{s} \neq \mathbf{0} \end{aligned} \tag{5.3}$$

The imposition of constraints $\mathbf{s} = \mathbf{0}$ is equivalent to assuming that g is a parameter, i.e., the functional form g is constant and independent from the experimental conditions φ. The constrained maximum likelihood estimate $\hat{\Theta}_m$ is obtained solving the constrained likelihood equations

$$\begin{aligned} \nabla\mathscr{L}_m(Y|\hat{\Theta}_m) + \nabla\mathbf{s}\hat{\boldsymbol{\alpha}} &= \mathbf{0} \\ \mathbf{s} &= \mathbf{0} \end{aligned} \tag{H.2}$$

where ∇ is the $(N + N_\theta - 1) \times 1$ gradient operator in the parameter space associated with Θ_m.

The Lagrange multipliers statistic is defined as

$$\xi_m(\theta_1) = \nabla\mathscr{L}_m(Y|\hat{\Theta}_m)^T \mathbf{H}_m^{-1} \nabla\mathscr{L}_m(Y|\hat{\Theta}_m) \sim \chi^2_{N-1} \tag{H.3}$$

where $\mathbf{H}_m$ is the $(N+N_\theta-1) \times (N+N_\theta-1)$ Fisher information matrix associated with the estimates $\hat{\Theta}_m$ (Bard, 1974).

The illustrated procedure for the construction of the statistic $\xi_m(\theta_1)$ can be repeated for diagnosing all the model parameters obtaining the set of statistics $\xi_m(\theta_i)\ \forall\, i = 1, ..., N$.

The multivariate MMI is then computed from each Lagrange multipliers statistic as

$$\mathrm{MMI}(\theta_i) = \frac{\xi_m(\theta_i)}{\chi^2_{N-1}(95\%)} \quad \forall\, i = 1,...,N_\theta \tag{H.4}$$

H.2 Case study and results

Model misspecification is diagnosed with the multivariate Lagrange multiplier in a simulated case study on a fed-batch bio-reactor system (Asprey and Macchietto, 2000). The case study is the same considered in Section 5.3.1. The system model involves the set of equations (5.10) and (5.11) with a Contois-type kinetic (5.12).

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = (r - u_1 - \theta_4)x_1 \tag{5.10}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \tag{5.11}$$

$$r = \frac{\theta_1 x_2}{\theta_2 x_1 + x_2} \tag{5.12}$$

The approximated model structure involves the same set of differential equations (5.10) and (5.11), but with a Monod-type kinetic (5.13).

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = (r - u_1 - \theta_4)x_1 \tag{5.10}$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \tag{5.11}$$

$$r = \frac{\theta_1 x_2}{\theta_2 + x_2} \tag{5.13}$$

Experimental data are generated in-silico by integrating the model equations with the system model. The parameter set assumed to simulate the experiments is $\theta^* = [0.310, 0.180, 0.550, 0.050]^T$. The same experimental design and measurement noise assumed in Section 5.3.1 are assumed here.

Table H.1: Comparison between the MMIs obtained neglecting parameter interaction and considering parameter interaction.

Parameter	Model Modification Indexes			
interaction	θ_1	θ_2	θ_3	θ_4
neglected	16.98	47.08	11.90	18.58
considered	52.74	53.39	16.05	43.07

The MMIs obtained neglecting parameter interaction (i.e., the MMIs computed with

the univariate Lagrange multipliers statistic ξ_d) and the MMIs obtained considering parameter interaction (i.e., the MMIs computed with the multivariate Lagrange multipliers statistic ξ_m) are reported in Table H.1. The MMIs are also reported in the radar charts in Figure H.1 for a visual comparison. In both cases, the MMIs associated with all the parameters are above unity. Hence, both the univariate MMIs and the multivariate MMIs suggest that a significant improvement in the model fitting quality is expected should any of the parameters be evolved.

As one can see, when parameter interaction is considered, all the MMIs increase in value. In particular, it is observed that the highest MMIs in the multivariate case are associated with parameters θ_2 and θ_1, respectively $\text{MMI}(\theta_2) = 53.39$ and $\text{MMI}(\theta_1) = 52.74$. These parameters are involved in the misspecified rate equation (5.13). It is observed that it is possible to make the approximated model indistinguishable from the system model either by evolving θ_2 into the function $\theta_2 x_1$, but also by evolving θ_1 into $\theta_1(\theta_2 + x_2)/(\theta_2 x_1 + x_2)$. Hence, the multivariate MMI-based analysis better highlights that also the evolution of θ_1 may result in a major improvement of the fitting quality.

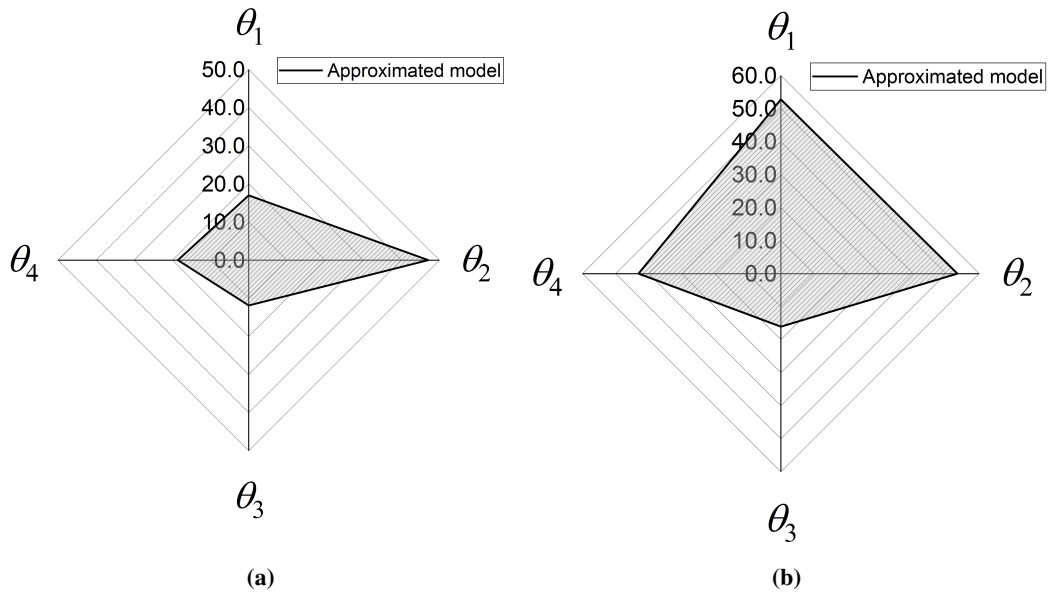


Figure H.1: Baker's yeast system. Model Modification Indexes associated with the approximated model (a) when the univariate Lagrange multipliers statistic ξ_d is used neglecting parameter interaction and (b) when the multivariate Lagrange multipliers statistic ξ_m is employed considering parameter interaction.

H.3 On the computability of the multivariate MMI

The $(N+N_\theta-1)\times(N+N_\theta-1)$ Fisher information matrix $\mathbf{H}_m$ used in the multivariate Lagrange multiplier test refers to the $(N+N_\theta-1)\times 1$ array of parameters $\Theta_m = [\theta_{1,1},...,\theta_{1,N},\theta_2,\theta_3,...,\theta_{N_\theta}]^T$. The information matrix is computed according to

$$\mathbf{H}_m = \sum_{i=1}^{N} \nabla\hat{\mathbf{y}}_i(\hat{\theta}_{1,i},\hat{\theta}_2,...,\hat{\theta}_{N_\theta})\Sigma_y^{-1}\nabla\hat{\mathbf{y}}_i(\hat{\theta}_{1,i},\hat{\theta}_2,...,\hat{\theta}_{N_\theta})^T \tag{H.5}$$

Matrix Σ_y is the covariance of the measurement noise associated to a $N_y\times 1$ sample of $\mathbf{y}$. A necessary condition for the computability of the Lagrange multipliers statistic ξ_m is that the matrix $\mathbf{H}_m$ is invertible. The matrix $\mathbf{H}_m$ can be decomposed in the following quadratic form

$$\mathbf{H}_m = \mathbf{Q}\begin{bmatrix} \Sigma_y & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_y & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_y \end{bmatrix}^{-1}\mathbf{Q}^T \tag{H.6}$$

Where $\mathbf{Q}$ is a $(N+N_\theta-1)\times N\cdot N_y$ sensitivity matrix constructed as follows

$$\mathbf{Q} = \begin{bmatrix} \frac{\partial y_{1,1}}{\partial\theta_{1,1}} & \cdots & \frac{\partial y_{1,N_y}}{\partial\theta_{1,1}} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial y_{2,1}}{\partial\theta_{1,2}} & \cdots & \frac{\partial y_{2,N_y}}{\partial\theta_{1,2}} & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \frac{\partial y_{N,1}}{\partial\theta_{1,N}} & \cdots & \frac{\partial y_{N,N_y}}{\partial\theta_{1,N}} \\ \frac{\partial y_{1,1}}{\partial\theta_2} & \cdots & \frac{\partial y_{1,N_y}}{\partial\theta_2} & \frac{\partial y_{2,1}}{\partial\theta_2} & \cdots & \frac{\partial y_{2,N_y}}{\partial\theta_2} & \cdots & \frac{\partial y_{N,1}}{\partial\theta_2} & \cdots & \frac{\partial y_{N,N_y}}{\partial\theta_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ \frac{\partial y_{1,1}}{\partial\theta_{N_\theta}} & \cdots & \frac{\partial y_{1,N_y}}{\partial\theta_{N_\theta}} & \frac{\partial y_{2,1}}{\partial\theta_{N_\theta}} & \cdots & \frac{\partial y_{2,N_y}}{\partial\theta_{N_\theta}} & \cdots & \frac{\partial y_{N,1}}{\partial\theta_{N_\theta}} & \cdots & \frac{\partial y_{N,N_y}}{\partial\theta_{N_\theta}} \end{bmatrix} \tag{H.7}$$

In (H.7), quantity $y_{i,j}$ represents the model prediction for the j-th measured response in the i-th sample. Since in (H.6), the matrix in the center of the quadratic form is positive definite, the rank of $\mathbf{H}$ is equal to the row-rank of the sensitivity matrix $\mathbf{Q}$. From this, it is possible to derive two necessary conditions for $\mathbf{Q}$ to be full row-rank:

- The number of columns of $\mathbf{Q}$ has to be greater or equal than the number of rows, i.e.,

$N \cdot N_y \geq N + N_\theta - 1$;

- The approximated model must be identifiable, i.e., the sensitivity matrix in the original parameter space Θ must have full row-rank N_θ.

If the latter condition is not satisfied, the row-rank of the $N_\theta \times N \cdot N_y$ sensitivity matrix in (H.8) is lower than N_θ. If the rows of the sensitivity matrix in (H.8) are linearly dependent, since $\hat{\theta}_{1,i} = \hat{\theta}_1 \ \forall\, i = 1, ..., N$, then also the rows in matrix $\mathbf{Q}$ are linearly dependent.

$$\begin{bmatrix} \frac{\partial y_{1,1}}{\partial \theta_1} & \cdots & \frac{\partial y_{1,N_y}}{\partial \theta_1} & \frac{\partial y_{2,1}}{\partial \theta_1} & \cdots & \frac{\partial y_{2,N_y}}{\partial \theta_1} & \cdots & \frac{\partial y_{N,1}}{\partial \theta_1} & \cdots & \frac{\partial y_{N,N_y}}{\partial \theta_1} \\ \frac{\partial y_{1,1}}{\partial \theta_2} & \cdots & \frac{\partial y_{1,N_y}}{\partial \theta_2} & \frac{\partial y_{2,1}}{\partial \theta_2} & \cdots & \frac{\partial y_{2,N_y}}{\partial \theta_2} & \cdots & \frac{\partial y_{N,1}}{\partial \theta_2} & \cdots & \frac{\partial y_{N,N_y}}{\partial \theta_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ \frac{\partial y_{1,1}}{\partial \theta_{N_\theta}} & \cdots & \frac{\partial y_{1,N_y}}{\partial \theta_{N_\theta}} & \frac{\partial y_{2,1}}{\partial \theta_{N_\theta}} & \cdots & \frac{\partial y_{2,N_y}}{\partial \theta_{N_\theta}} & \cdots & \frac{\partial y_{N,1}}{\partial \theta_{N_\theta}} & \cdots & \frac{\partial y_{N,N_y}}{\partial \theta_{N_\theta}} \end{bmatrix} \tag{H.8}$$

Future research activities may focus on the identification of sufficient conditions for matrix $\mathbf{Q}$ to be full row-rank. It is recognised that an experimental design approach may be employed to design samples with the purpose of obtaining a non-singular information matrix $\mathbf{H}_m$.