UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN CHEMICAL AND PROCESS ENGINEERING

Tesi di Laurea Magistrale in

Chemical and Process Engineering

Using estimability analysis to rank and select model parameters

Relatore: Prof. Fabrizio Bezzo
Correlatrice: Ing. Geremia Margherita

Laureando: Andrea Valsasina

ANNO ACCADEMICO 2023/2024

Abstract

Mathematical models can be used to support the development and optimization of chemical processes. However, modelers often face challenges in estimating all the model parameters due to identifiability and estimability issues. This is typically the case of complex and/or over-parameterized models, resulting in some parameters being unidentifiable or difficult to be precisely estimated. Moreover, data are usually limited, while running additional experiments might be expensive or impractical. To tackle this issue, modelers often estimate only a subset of all the model parameters.

In this work, a systematic workflow for parameter estimability proposed in the literature (Wu et al., 2011, *International Journal of Advanced Mechatronic Systems*, **3**, pp. 188-197) is implemented and critically evaluated. The methodology is applied to two case studies of industrial interest: (i) a fermentation process, and (ii) a process for the production of urethane. Different levels of uncertainty on parameter values and experimental noise are taken into account. A robustness test is used to assess the sensitivity of the methodology to initial parameter guesses. Results suggests that parameters ranking and subset selection depends significantly on both the initial guesses of the model parameters and sensors noise, and that repeating the methodology can improve the results and the precision of parameter estimates.

Table of contents

List of symbols

Acronyms

ARE	=	average relative error
CAPE	=	computer-aided process engineering
CI	=	confidence interval
DAE	=	differential and algebraic equations
EM	=	extended model
FIM	=	Fisher information matrix
GSA	=	global sensitivity analysis
LSA	=	local sensitivity analysis
MBDoE	=	model-based design of experiments
MC	=	Monte Carlo
MSE	=	mean squared error
ODE	=	ordinary differential equation
OLS	=	ordinary least squares
SM	=	simplified model
SQP	=	sequential quadratic programming
SSE	=	sum of squared errors

Greek letters

α	=	significance level
Δh	=	reaction enthalpy [J/mol]
$\boldsymbol{\varepsilon}$	=	vector of additive random noise terms
$\boldsymbol{\theta}$	=	vector of model parameters
θ_{j0}	=	initial guess value of parameter θ_j
v_{ii}	=	iith term of the variance-covariance matrix
ρ_i	=	species densities [kg/m^3]
σ	=	measurement noise standard error
σ^2	=	measurement noise variance
Σ^{-1}	=	inverse of the variance-covariance matrix
τ	=	total duration of the experiment
$\boldsymbol{\varphi}$	=	experiment design vector
$\boldsymbol{\varphi}^{\mathrm{opt}}$	=	optimal experimental design vector

Latin letters

c_i	=	species concentration [mol/L]
d	=	number of response variables
e	=	stochastic component
E_a	=	activation energy [J/mol]
f	=	vector of model equations
f^{v1}	=	accumulated feed of vessel 1 [mol]
f^{v2}	=	accumulated feed of vessel 2 [mol]
h	=	vector of model equations
J	=	parameter estimation objective function
J_k	=	objective function when k parameters are estimated
J_p	=	objective function when all parameters are estimated
k_{c2}	=	equilibrium constant [L/mol]
k_{ref}	=	steric factor [L/mol · h]
$lb_{\theta j}$	=	lower bound value of parameter θ_j
M_i	=	species molar masses [kg/mol]
n	=	number of measurement times
n_i	=	species molar number [mol]
n_i^0	=	species initial molar number [mol]
n_i^{v1}	=	species molar number on vessel 1 [mol]
n_i^{v2}	=	species molar number on vessel 2 [mol]
p	=	number of parameters
p_1	=	number of parameters in the SM
r	=	number of experimental runs
R	=	gas constant [J/(mol · K)]
R^2	=	coefficient of determination
r_C	=	critical ratio estimator
R_C	=	critical ratio
r_{CC}	=	corrected critical ratio estimator
R_{CC}	=	corrected critical ratio
$r_{CC,k}$	=	corrected critical ratio estimator when k parameters are estimated
$r_{C,Kub}$	=	Kubokawa critical ratio estimator
$r_{C,Kub,k}$	=	Kubokawa critical ratio estimator when k parameters are estimated
$\mathbf{R}_k$	=	residual matrix with k model parameters evaluated
S	=	sensitivity matrix
S_i	=	first-order index of Sobol analysis

S_{ij}	=	sensitivity coefficient
$S_{T,i}$	=	total-effect index of Sobol analysis
$s_{\theta j0}$	=	scaling factor representing the uncertainty on parameter values
s_{yi}	=	scaling factor representing the uncertainty on the measurements
t	=	time [h]
T	=	temperature [K]
T_{g2}	=	reference temperature [K]
t_i	=	t-value of the ith model parameter
t_{ref}	=	reference t-value
T_{ref}	=	reference temperature [K]
$\mathbf{t}_{\mathrm{sp}}$	=	vector of output variables sampling times
$ub_{\theta j}$	=	upper bound value of parameter θ_j
u_1	=	dilution factor [h^{-1}]
u_2	=	substrate concentration in the feed [g/L]
$\mathbf{u}(t)$	=	vector of time varying control input variables
V	=	volume [L]
$V(y_j)$	=	variance of the jth model output
$V_{12\ldots k}$	=	kth-order contribution of factor i to the variance of the output
V_i	=	first-order contribution of factor i to the variance of the output
V_{ij}	=	second-order contribution of factor i to the variance of the output
V_{θ_i}	=	variance based first order effect for factor θ_i
$V_{\theta\sim i}$	=	variance based first order effect for factor $\theta_{\sim i}$
$\mathbf{V}_\theta$	=	parameter variance-covariance matrix
$\mathbf{V}_\theta^0$	=	preliminary variance-covariance matrix
$\mathbf{x}(t)$	=	vector of time dependent state variables
$\dot{\mathbf{x}}(t)$	=	vector of differential state variables
$\mathbf{X}_\mathrm{k}$	=	scaled sensitivity matrix of the kth iteration
x_1	=	biomass concentration [g/L]
x_1^0	=	initial biomass concentration
x_2	=	substrate concentration [g/L]
x_2^0	=	initial substrate concentration
$\mathbf{y}(t)$	=	vector of measured outputs
$\hat{\mathbf{y}}(t)$	=	vector of outputs predicted by the model
$\mathbf{y}_0$	=	vector of state variables initial conditions
$\mathbf{Z}$	=	scaled sensitivity matrix
$\hat{\mathbf{Z}}_\mathrm{k}$	=	Predicted scaled sensitivity matrix of the kth iteration

Introduction

The accurate estimation of parameters in mathematical models is essential for describing chemical processes and predicting their behavior under various conditions. However, this task often poses significant challenges due to the inherent complexity, non-linearity, and overparameterization of many chemical process models. As the number of parameters increases, so does the potential for issues such as parameter correlation, non-identifiability, and lack of convergence in estimation algorithms. These challenges are further compounded by the nature of the experimental data, which are often sparse, noisy, or uninformative. In many cases, performing additional experiments to gather more informative data is either cost-prohibitive or technically unfeasible.

In light of these difficulties, analyzing the identifiability and estimability of model parameters becomes a critical step before parameter estimation can proceed.

In Chapter 1 the context of modeling when chemical processes are involved is presented, underlying the difficulties that modelers may face when model parameters need to be estimated. Following a review of existing methods in the literature that address issues related to parameter identifiability and estimability, motivations and objectives associated to a more rigorous and novel estimability method are discussed.

Chapter 2 describes the estimability method proposed by Wu et al., (2011) and the orthogonalization algorithm – which is required to produce a ranked list of model parameters – as well as the mathematical techniques adopted to implement the method algorithms. Additionally, two other estimability methods based on sensitivity analysis – i.e., local sensitivity analysis (LSA), and global sensitivity analysis (GSA) – are introduced for comparative purposes. Finally, a technique for evaluating the robustness of the results is presented.

The validity of the estimability method is then assessed in Chapter 3, where a fermentation process is used as a case study. The method is applied under varying levels of measurement noise and uncertainty on parameter values. Robustness of these results is then examined, and the precision of parameter estimates is analyzed. Finally, the results are compared with those obtained using the sensitivity-based methods discussed in the previous chapter.

In Chapter 4, the estimability method is applied to a more complex model describing the production of urethane. Results are then commented, and similarities with the previous case study are discussed, together with the precision of parameter estimates and model performance. Lastly, the robustness of the results is examined.

Some final remarks conclude the work.

Chapter 1

Assessing parameter estimability in mathematical models

This chapter addresses current challenges associated with the implementation of mathematical process models, with a particular focus on difficulties in estimating all model parameters based on available experimental data. Methods proposed in literature – which are aimed at determining whether and which parameters in a model can be estimated from the experimental data – are presented and analyzed. Finally, the motivations and objectives of this work are discussed, with particular focus on a recently proposed approach retrieved from literature.

1.1. Challenges on mathematical modelling of chemical processes

Over the last decades, the improvement of computational power led to an increase in the development of computational methods that address engineering challenges. With that, the field of Computer-Aided Process Engineering (CAPE) is dramatically transforming the way industrial processes are developed, designed and optimized (Quaglio, 2020).

The CAPE approach involves implementing mathematical models of processes into computer programs. Mathematical models are valuable tools for developing new chemical processes and enhancing existing ones. These models enable chemical engineers to gain a deeper understanding of the actual operating process and predict its behaviour under various operating conditions. Furthermore, they are instrumental in training operators and in the development and testing of control systems (Foss et al., 1998).

These models are generally formulated as systems of differential and algebraic equations, with their mathematical structure representing the underlying causal mechanisms of the physical system, and are assumed to take the following general form:

$$\mathbf{f}(\dot{\mathbf{x}}(t), \mathbf{x}(t), \mathbf{u}(t), t, \boldsymbol{\theta}) = 0 \quad , \tag{1.1}$$

$$\hat{\mathbf{y}} = \mathbf{h}(\mathbf{x}(t), \mathbf{u}(t), t, \boldsymbol{\theta}) \quad , \tag{1.2}$$

where $\mathbf{f}$ and $\mathbf{h}$ represent model equations, $\mathbf{u}$(t) represent control input variables, t is time, $\boldsymbol{\theta}$ represents model parameters and $\hat{\mathbf{y}}$ represents model predictions for a measurable set of system states $\mathbf{y}$.

In practice, developing fundamental models presents several challenges, as the processes of interest to chemical engineers are often highly complex and involve numerous reactions with

associated kinetic parameters. The complexity of these models further increases when thermodynamic and mass transfer parameters are incorporated to more accurately describe the process.

As shown in Figure 1.1, several steps are used to develop fundamental models. Parameter estimation (Step 4) contains its set of challenges, since models are generally nonlinear in the parameters, requiring nonlinear optimization algorithms and initial guesses of parameter values for parameter estimation. Initial guesses of parameter values become then very important because multiple optima may be encountered. Additionally, most of the times modelers frequently work with limited and/or noisy experimental data, making it nearly impossible to accurately estimate all model parameters. Some parameters may have minimal impact on model predictions, rendering precise estimation unattainable. In other cases, the effects of certain parameters on model predictions may be correlated with those of others, leading to different sets of parameter values that yield nearly identical predictions (Petersen et al., 2001).

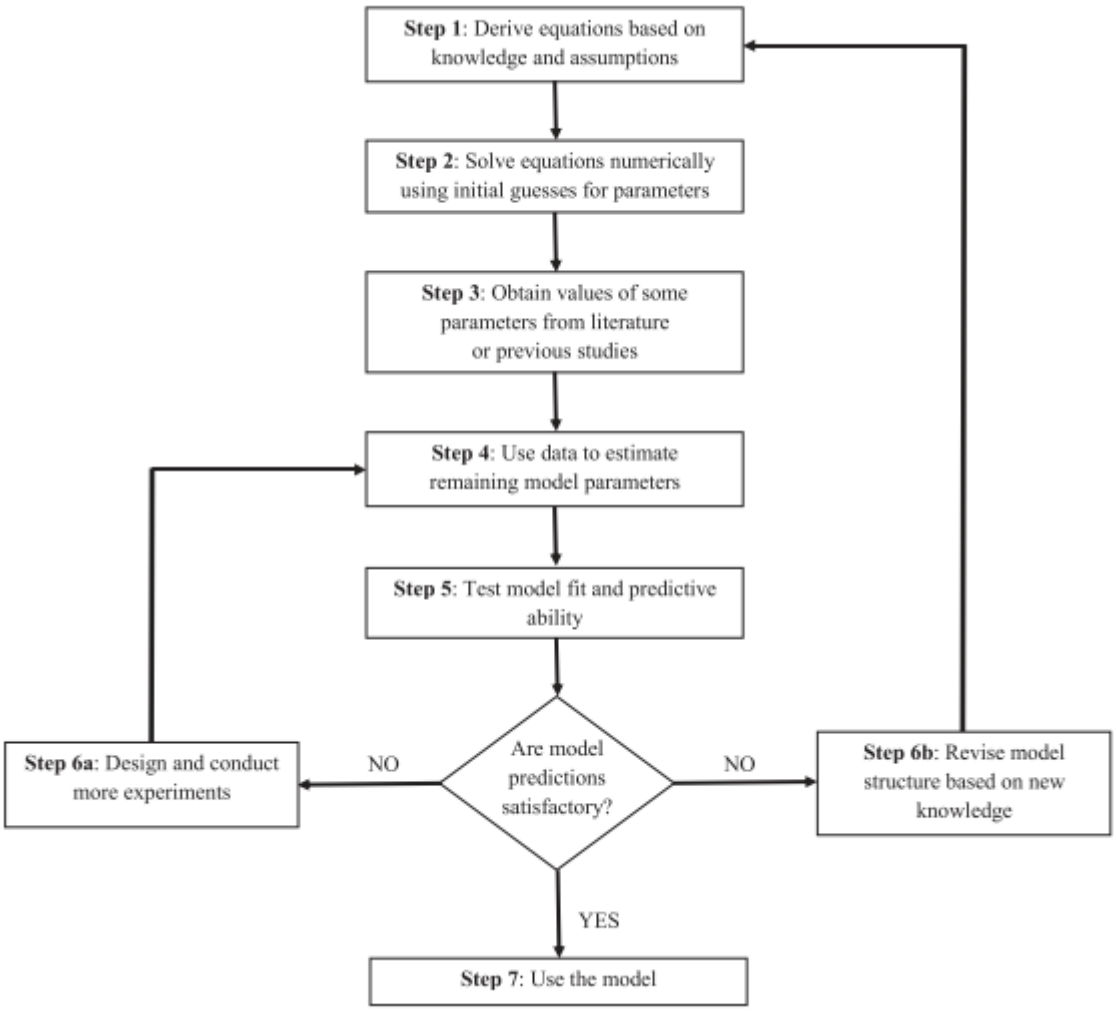


Figure 1.1. *Steps in model formulation (McLean and McAuley, 2012).*

In this context, it may be useful to check the identifiability and estimability of model parameters. Identifiability issues arise when the model structure is the main problem when the parameters cannot be uniquely estimated, while estimability issues arise when it is not possible to obtain unique parameter estimates with the existing experimental data. Table 1.1 summarizes the concepts of parameter identifiability and estimabilty.

Table 1.1. *Identifiability and estimability (McLean and McAuley, 2012).*

	Identifiability analysis	**Estimability analysis**
Question answered	Can different values of model parameters lead to the same input-output behaviour for the model?	Can all parameter values be estimated uniquely from the available experimental data?
Alternative names	Structural identifiability A-priori identifiability	Practical identifiability A-posteriori identifiability
Model complexity	Difficult to assess in models with more than ~10 parameters	Can be assessed in models with ~50 parameters

Sometimes model parameters are identifiable based on the model structure but may not be estimable in practice due to limited information contained in the available experimental data. In these cases, fixing some parameters at their nominal values and estimating only a subset of them can make possible the parameter estimation task.
Techniques used to assess the identifiability and estimability of model parameters will be discussed next.

1.2. Identifiability Analysis

Identifiability analysis, also called structural identifiability and/or a-priori identifiability, can be performed to determine if unique values of model parameters can be obtained under the assumptions of accurate model structure and noise-free measurement data (Kim and Lee, 2019). A model and its parameters are said to be identifiable if there exists a unique input-output behaviour for each set of parameter values. Consider a nonlinear ordinary differential equation (ODE) model of the form:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}) \quad , \tag{1.3}$$

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}) \quad , \tag{1.4}$$

where **x** is a vector of model states, **u** is a vector of input trajectories, **θ** is a vector of model parameters and **y** is a vector of model predictions. The model and parameters are said to be identifiable if the following holds for any θ_1 and θ_2 belonging to the allowable parameter space:

$$\mathbf{g}(\mathbf{x}, \mathbf{u}, \theta_1) = \mathbf{g}(\mathbf{x}, \mathbf{u}, \theta_2) \Leftrightarrow \theta_1 = \theta_2 \quad , \tag{1.5}$$

if Equation (1.5) holds for the entire parameter space, the parameters are said to be globally identifiable, whereas if Equation (1.5) holds only for a neighbourhood of the parameter space, the parameters are said to be locally identifiable (Jimènez-Hornero et al., 2008).

Assessing identifiability can uncover problems with model structure, and if model parameters are unidentifiable, the modeller will need to either simplify the model equations or obtain more information from additional output variables. Performing additional experiments using the same measured outputs will not solve an identifiability problem, therefore, the modeller could simplify the model by either lumping several parameters that appear together in one single overall parameter (e.g., Chu and Hahn, 2009), or remove terms that are expected to have little influence on model predictions (e.g., Degenring et al., 2004), or fix some parameters, thus reducing the number of parameters that require estimation (e.g., Chu et al., 2009).

Several methods were developed to assess identifiability in nonlinear ODE models. Grewal and Glover (1976) proved that if the linearised version of a nonlinear model is identifiable, then the original nonlinear model is also identifiable. Unfortunately, if the linearised version of the nonlinear model is unidentifiable, we cannot conclude that the original nonlinear model is also unidentifiable (Vajda et al., 1989).

Pohjanpalo (1978) proposed a Taylor-series-expansion approach, which require high order derivatives of system outputs with respect to time, where the number of derivatives required is at least equal to the number of parameters. Vajda and Rabitz (1989) developed a similarity transformation (or local state isomorphism) approach, Ljung and Glad (1994) developed a differential algebra technique that is restricted to polynomial nonlinearities. The latter method has been applied using symbolic computational packages, eliminating the need to perform complicated algebraic equations by hand. Jimenez-Hornero et al. (2008) attempted to use the cited techniques to assess the identifiability of a nonlinear ODE model of an acetic acid fermentation process. All the cited techniques either failed to provide firm conclusions or were too difficult to implement due to the complexity of the analytical expressions and long computational times.

In summary, even though structural identifiability represents a “low bar” in order to a model be identifiable, structural identifiability analysis methods are not widely used in practice, yet, due to either the computational complexity or the lack of mature computer implementations (Miao et al., 2011).

1.3. Estimability Analysis

Parameters that are said to be identifiable, may not be estimable in practice due to the limited information contained in the available experimental data. Estimability analysis, also known as practical identifiability and/or a-posteriori identifiability, is a tool for determining if all model parameters can be estimable from the available data.

An easier and often better choice is to first assess the estimability of the model parameters, since many identifiability techniques are complicated and can fail when applied to nonlinear fundamental models. Techniques for determining estimability are easier to use and can be applied to complex nonlinear models with many parameters (McLean and McAuley, 2012).

A parameter may be non-estimable for two primary reasons: (1) the model predictions are not sensitive to variations in the parameter's value, or (2) the parameter's influence on model predictions is correlated with the influence of one or more other parameters. Assessing parameter estimability requires evaluating both the sensitivity of model predictions to changes in parameter values and the correlations between the effects of different parameters. The methods for assessing parameter estimability will be discussed in greater detail.

1.3.1. Sensitivity-based estimability methods

Sensitivity analysis is a well-researched and utilized concept that can be often used to rank the influence of parameters on the model outputs through calculations of sensitivity matrices and can in some cases determine non-estimability. The associated sensitivities of the predicted outputs to the parameters can be placed in a sensitivity matrix, **S**:

$$\mathbf{S} = \begin{pmatrix} \left.\frac{\partial \hat{y}_{11}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{11}}{\partial \theta_p}\right|_{u_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{1n}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{1n}}{\partial \theta_1}\right|_{u_1} \\ \left.\frac{\partial \hat{y}_{21}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{21}}{\partial \theta_p}\right|_{u_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_p}\right|_{u_1} \\ \left.\frac{\partial \hat{y}_{11}}{\partial \theta_1}\right|_{u_2} & \cdots & \left.\frac{\partial \hat{y}_{11}}{\partial \theta_p}\right|_{u_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_1}\right|_{u_2} & \cdots & \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_p}\right|_{u_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_1}\right|_{u_r} & \cdots & \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_p}\right|_{u_r} \end{pmatrix} , \quad (1.6)$$

where the dimensions of **S** are $[N\ X\ p]$, with $N=dnr$, where d is the number of response variables, n is the number of measurement times, r is the number of experimental runs and p is the number of parameters. Each column corresponds to the sensitivities of the predicted outputs of a single parameter.

For simple models with a limited number of parameters and outputs, graphical inspection of the sensitivity matrix can provide insights into the estimability of the parameters. For instance, visual analysis of the sensitivity matrix can reveal whether model predictions are insensitive to variations in a parameter or whether combinations of two or more parameters produce similar model outputs, indicating a correlation among these parameters (Holmberg, 1982). Both scenarios can lead to issues with parameter estimability. For example, Figure 1.2 shows that parameters θ_2 and θ_3 have little to no influence on the model output, while parameters θ_1 and θ_4 are shown to have great influence and their effect on the model output seem to be correlated with each other. It is crucial to ensure appropriate scaling when calculating sensitivity coefficients to prevent small sensitivity values from arising due to an improper choice of units for the parameters or model outputs.

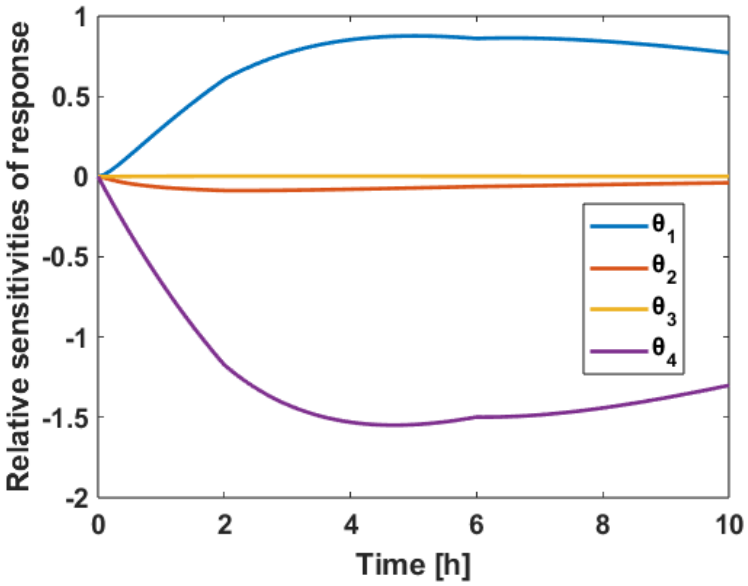


Figure 1.2. *Sensitivity plot.*

Other methods that rely on the sensitivity matrix are based on the Fisher Information Matrix (FIM). The FIM is important for estimability analysis since it contains information about measurement uncertainty. The sensitivity matrix in equation (1.4) is related to the FIM by:

$$FIM = \mathbf{S}^{\mathrm{T}}\Sigma^{-1}\mathbf{S} \quad , \tag{1.7}$$

where Σ^{-1} is the inverse of the variance-covariance matrix of the response variables. A simple way to assess estimability is to check the rank of the FIM. If the rank of the FIM is less than the number of the parameters (i.e. the FIM is singular) then the parameters are not all estimable (Petersen et al. 2001). Due to numerical uncertainties, it is difficult in many times to determine if the FIM has full rank or not, which is why in many cases the FIM is analyzed via an eigen-decomposition. The presence of a "small" eigenvalue is indicative of ill-conditioning of the matrix, and as a result the corresponding parameter cannot be estimated reliably (Lam et al., 2022).

In summary, sensitivity-based estimability methods are rather simple to implement but in practice are recommended for simple models with few parameters. Moreover, these methods need to be implemented carefully, since the calculations of the sensitivity matrix are based on initial parameter guesses that may be inaccurate, and conclusions are valid locally, (i.e. in the proximities of the initial parameter guess).

1.3.2. Global sensitivity analysis

Global sensitivity analysis (GSA) can be used to quantify the importance of model parameters and their interaction with respect to model outputs. Differently from a local sensitivity analysis, the results of a GSA do not depend on a specific choice of parameters value, but it uses as input the entire range of parameters values (Braakman et al., 2022).
One of the most used GSA methods, also used during this project, is the variance-based Sobol's GSA. This is a variance-based methodology based on Saltelli (2002) in which the variance of the ith model output is decomposed as follow:

$$V(y_j) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \cdots + V_{12\ldots k} \quad , \tag{1.8}$$

where V_i is the first-order effect of factor i to the variance of the output, while V_{ij} and $V_{12\ldots k}$ express the variance contributions due to second and k-th order interactions. There are two types of variance-based sensitivity indices (Sobol, 1993): the *first-order index* (S_i) ,which represents the main effect contribution of each input factor (in our context, the parameters) to the variance of the output. It indicates by how much one could reduce, on average, the output variance if the corresponding input factor could be fixed. The other index is the *total effect index* ($S_{T,i}$) that accounts for the total contribution to the output variance of the i-th factor, including its individual contribution (first-order effect) plus all higher-order effects due to its interactions with other factors. The first-order index is defined as:

$$S_i = \frac{V_{\theta_i}(E_{\theta\sim i}(y_j|\theta_i))}{V(y_j)} \quad , \tag{1.9}$$

where $V_{\theta_i}(E_{\theta\sim i}(y_j|\theta_i))$ represents the first-order effect of θ_i on the output. The higher the value of S_i is, the higher the influence of the i-th factor on the output. The total effect index is defined as:

$$S_{T,i} = 1 - \frac{V_{\theta\sim i}\left(E_{\theta_i}(y_j|\boldsymbol{\theta}_{\sim \mathrm{i}})\right)}{V(y_j)} \quad , \tag{1.10}$$

where $V_{\theta\sim i}(E_{\theta_i}(y_j|\boldsymbol{\theta}_{\sim \mathrm{i}}))$ is the first-order effect of $\boldsymbol{\theta}_{\sim \mathrm{i}}$. If the value of $S_{T,i}$ is small, it is a necessary and sufficient condition for θ_i being noninfluential. Interpretations about Sobol's GSA can be summarized as follows:

- The higher the value of S_i, the higher the influence of the i-th factor on the output;
- If $S_i = 0$, then the i-th factor has no direct influence on the output; however, it may still be an important factor through its interactions with other factors;

- A significant difference between S_i and $S_{T,i}$ indicates an important interaction involving that factor;
- The sum of all S_i is always lower than or equal to 1. If it is equal to 1, then there are no interactions between the factors and this implies that the model is additive;
- $S_{T,i}$ must be higher or equal to S_i. If it is equal, then the factor has no interactions with the other factors;
- If $S_{T,i} = 0$, the i-th factor has no influence on the model output and the factor can be fixed at any value within its range of uncertainty;
- The sum of all $S_{T,i}$ is always higher than or equal to 1. If it is equal to 1, then there are no interactions between the factors.

Sobol's Global Sensitivity Analysis (GSA) method can be effectively employed to assess parameter estimability, as it identifies parameters that have little to no influence on model outputs and detects significant interactions between parameters across the entire range specified by the user. This makes Sobol's GSA a valuable tool for reducing the number of "free" parameters in a model by fixing those that are non-influential.

1.3.3. Repeated parameter estimation methods

Monte Carlo (MC) simulations can be used to assess parameter estimability. The MC method used for this purpose (Miao et al., 2011) repeatedly simulates measurement noise on a dataset. The general procedure of the method is as follows:

1. Determine a nominal parameter vector $\widehat{\boldsymbol{\theta}}$ for simulation study.
2. Simulate the model with $\widehat{\boldsymbol{\theta}}$ to obtain measurements $\boldsymbol{y}$ at the experimental sample times.
3. Generate n (e.g., 1000) sets of simulated data from the measurements with a given measurement error level.
4. Fit the model to each of the n simulated datasets to obtain parameter estimates $\boldsymbol{\theta}_i$, i=1,2,…,n.
5. Calculate the average relative estimation error (ARE) for each element of $\boldsymbol{\theta}$ as

$$ARE\left(\theta^{(k)}\right) = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{\theta}^{(k)} - \theta_i^{(k)}\right|}{\left|\hat{\theta}^{(k)}\right|} \quad , \tag{1.11}$$

where $\hat{\theta}^{(k)}$ is the kth element of $\widehat{\boldsymbol{\theta}}$ and $\theta_i^{(k)}$ is the kth element of $\boldsymbol{\theta}_i$.

Steps 2-5 may be repeated for different levels of measurement noise, to understand how parameter estimability varies with measurement noise. ARE values are typically used to assess parameter estimability. For an expected level of measurement noise, high values of ARE indicate that a parameter is not estimable.

While this method is straightforward to apply, it is computationally expensive due to the large number of parameter estimations required. Additionally, it may produce misleading results if the parameter values used to generate the dataset are significantly different from the "true" parameter values.

Raue et al. (2009) developed a graphical profile-likelihood method to assess parameter estimability. Numerous parameter estimation calculations are required and the importance of each parameter on the objective function J used for parameter estimation is considered. The method consists in choosing a range of values for each parameter of interest, and for different candidate values from the range, the remaining parameters are adjusted to minimise the value of J. If the parameter of interest has a negligible effect on J (e.g., θ_2 in figure 1.3), the parameter is inestimable, and if the value of J is strongly dependent on the value of the parameter of interest (e.g., θ_1 in figure 1.3), the parameter is estimable.

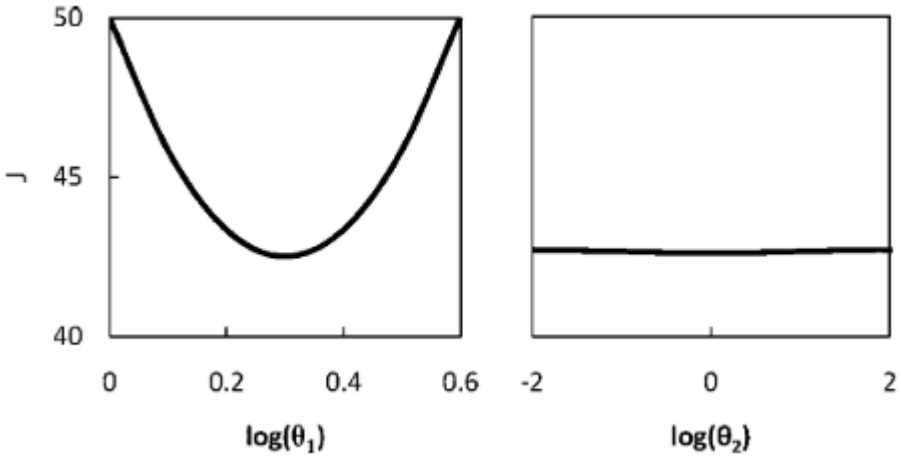


Figure 1.3. *Profile-likelihood estimability plots (Raue et al., 2009).*

Both cited methods require a threshold value (e.g., for the ARE and for changes in J to categorise parameters as estimable or inestimable).

All the estimability methods discussed can be applied to identify parameters in a model that may cause estimability issues. However, in practice, mathematical models of chemical and pharmaceutical processes are often highly complex and involve a large number of parameters, making it likely that some parameters will be unidentifiable or inestimable. In such cases, it may be practical to fix certain parameters at nominal values and estimate only a subset of them. A methodology to address this challenge will be presented in the next chapter.

1.4. Motivations and objectives

Over the last decades, significant efforts have been devoted in developing methods and techniques that can assist parameter estimation by determining whether and which model parameters can be estimated from the available experimental data. These methods, including identifiability and estimability approaches, however, pose some issues.

- Their implementation can be computationally demanding;
- Results may only be valid locally, e.g., in the proximity of a specific parameter;
- Outcomes can be influenced by arbitrary decisions and cut-off values, which are required to assess estimability.

Recently, the estimability method proposed by Wu et al. (2011) has gained attention due to its simplicity of implementation. It can be applied to models with numerous parameters to determine which parameters should be estimated from a given dataset.

This thesis aims to evaluate the recent estimability method by Wu et al. (2011), in combination with the well-known parameter ranking method proposed by Yao et al. (2003). The objective is to apply the estimability method to different models and compare it with other established approaches presented in literature. The method will then be analyzed under varying levels of measurement noise and uncertainty on parameter values, both of which can significantly influence the results.

Chapter 2

Methodology

In this chapter, a discussion about software tools and mathematical techniques used in this work is provided. Then, the methods used for ranking and selecting model parameters are explained thoroughly, and their algorithms are presented. At the end, the interest is focused on how these methods were implemented in this study, and on how the robustness of the results can be tested.

2.1. Software and mathematical tools

The project development requires a software capable of solving the system of algebraic and ordinary differential equations (DAEs) describing the process of interest. In this study, MATLAB® R2023b was used to solve the systems of DAEs, to implement the algorithms for parameter ranking and subset selection – which will be presented in the following –, and finally, to perform parameter estimation. It was also used Siemens Process System Enterprise's advanced process modelling software gPROMS® Model Builder v. 7.0. for performing global sensitivity analysis (GSA) and model-based design of experiments (MBDoE) to design informative experiments for the subsequent parameter estimation activity.

2.1.1. Model simulation

In this study, the mathematical models on which the parameters ranking and selection is assessed, are comprised by systems of ordinary differential equations and algebraic equations. In order to solve them, the equations of the model were implemented in MATLAB® R2023b, where the ODE solver `ode15s` was used. This solver is able to solve stiff differential equations and DAEs, requiring as input the time span of the integration, an initial value of the response variables and the values of the eventual manipulated variables and parameters.

2.1.2. Parameter estimation

Once the mathematical model and the experimental data are available, parameter estimation is performed. The aim is to estimate the values of model parameters that allow to obtain the model predictions as close as possible to the observed experimental data. This is achieved by minimizing the value of an objective function (Braakman et al., 2022). In this study, as requested by the parameter selection method, explained in detail in §2.4, parameter estimation is performed by minimizing the weighted sum of squared residuals of the following objective function J:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{d} \sum_{l=1}^{n} \sum_{m=1}^{r} \left(\frac{y_{ilm} - \hat{y}_{ilm}(\boldsymbol{\theta})}{s_{yi}} \right)^2 , \tag{2.1}$$

where y and $\hat{y}$ are the response and predicted response of the $i = 1 \dots d$ response variables, of $l = 1 \dots n$ measurement times and $m = 1 \dots r$ experimental runs. In this study, parameter estimation, and therefore the minimization of the objective function of Equation (2.1), was performed by using a nonlinear programming solver on MATLAB® R2023b called `fmincon`, which is often used to find the minimum of constrained nonlinear multivariable functions. The solver needs as inputs the objective function, an initial guess of the parameters to be estimated and a constraint on the parameters, as a lower bound and an upper bound values. Setting those bound helps the solver to converge on meaningful parameter values. Additionally, a sequential quadratic programming algorithm (SQP) is used as algorithm for the solver, since SQP methods represent the state of the art in nonlinear programming methods and the algorithm is sometimes faster and more accurate than the default one (Schittkowski, 1986).

The parameter estimation task can be also very challenging considering that models often have many parameters to be estimated, and the data used to estimate those parameters are often sparse. This combination of sparse data and large number of parameters, along with nonlinearities in the model, can result in the optimization algorithm to converge in a local minimum of the objective function, yielding misleading parameter estimates (Braakman et al., 2022). Unless knowledge of the initial estimates is well known, it is recommended to use global optimization algorithms, in order to reduce the likelihood of ending up in a local minimum.

Taking this into account, parameter estimation in this study was performed by enabling a `MultiStart` option in the solver. This allows the solver to find multiple local minima, starting the solver from multiple starting points, or better, from multiple initial parameter values inside the constraints, so that the final local minimum picked by the solver will be the optimal between all the local minimum encountered.

2.1.3. Sensitivity analysis

Sensitivity analysis can be utilized to quantify the extent to which changes in various parameters affect the model outputs, and as seen in §1.3.1 and §1.3.2, it can also be used to assess parameter estimability. Sensitivity analysis provides metrics to rank the importance of parameters with respect to an output of interest: the larger the values of the metrics, the more sensitive the model response with respect to parameter changes (Saltelli et al., 2008). Parameters that are less influential therefore can be fixed to their nominal values and be excluded from the parameter estimation task, reducing the number of free parameters in the model. This reduction of free parameters can provide greater confidence in model output predictions (Braakman et al., 2022).

Generally, there are two types of sensitivity analysis:

- Local sensitivity analysis (LSA): performed with a one-at-a-time (OAT) method. A single parameter is perturbed around its nominal value, while the others remain fixed, and the size of the changes in the responses is noted, quantified by means of partial derivatives;
- Global sensitivity analysis (GSA): quantify the sensitivity of the model responses to the model parameters exploring the entire interval given for each parameter.

In this study, both analyses were used. LSA analysis is very straightforward since the partial derivatives of the model responses with respect to parameter changes are already calculated for the method that will be presented in §2.3. While for the GSA, the technique adopted was the Sobol method, already presented in §1.3.2, a variance-based methodology based on Saltelli (2002) in which the variance of the ith model output is decomposed as follow:

$$V(y_j) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \cdots + V_{12\ldots k} \ , \tag{2.2}$$

where V_i is the first-order effect of factor i to the variance of the output, while V_{ij} and $V_{12\ldots k}$ express the variance contributions due to second and k-th order interactions. The metrics used to quantify the influence of model parameters to the model responses are the first order index S_i and the total effect index $S_{T,i}$. The first order index is defined as:

$$S_i = \frac{V_{\theta_i}(E_{\theta\sim i}(y_j|\theta_i))}{V(y_j)} \ , \tag{2.3}$$

where $V_{\theta_i}(E_{\theta\sim i}(y_j|\theta_i))$ represents the first-order effect of θ_i on the output. The first order effect S_i represents the main effect of each parameter on the variance of the output. If a first-order index is large, then the corresponding parameter is influential on the output. The total effect index is defined as:

$$S_{T,i} = 1 - \frac{V_{\theta\sim i}\left(E_{\theta_i}(y_j|\boldsymbol{\theta}_{\sim \mathrm{i}})\right)}{V(y_j)} \ , \tag{2.4}$$

where $V_{\theta\sim i}(E_{\theta_i}(y_j|\boldsymbol{\theta}_{\sim \mathrm{i}}))$ is the first-order effect of $\boldsymbol{\theta}_{\sim \mathrm{i}}$. The total effect index accounts for the total contribution to the output variance of the i-th factor. If a total effect index is small, then the corresponding parameter is non-influential. Properties of the sensitivity indices are listed in §1.3.2.

In this work, GSA was performed within gPROMS® through the `Global System Analysis` entity that allows investigating the system behaviour. Variability of selected outputs is computed after specifying the range of variability of each input of interest.

2.1.4. Model based design of experiments

Parameters are estimated based on data obtained from experiments; therefore, experiments should be carefully designed to support effective parameter estimation. Model based design of experiments (MBDoE) techniques are effective mathematical tools, which exploit the knowledge of the structure of the underlying system, represented by a mathematical model, aiming to obtain the maximum information content from the designed experiment. These techniques are used to reduce the parameter uncertainty region through the optimization of the experiment design vector $\boldsymbol{\varphi}$, defined as (Franceschini and Macchietto, 2008):

$$\boldsymbol{\varphi} = [\mathbf{y}_0, \mathbf{u}(t), \mathbf{t}^{\mathrm{sp}}, \tau]^T \quad , \tag{2.5}$$

where $\mathbf{y}_0$ is the set of initial conditions for the measured variables, $\mathbf{u}(\mathrm{t})$ is the vector of time varying control variables, $\mathbf{t}^{\mathrm{sp}}$ is the vector of the output variables sampling times, and τ is the total duration of the experiment. The optimization problem is then formulated as:

$$\boldsymbol{\varphi}^{\mathrm{opt}} = \underset{\boldsymbol{\varphi}}{argmin}\{\psi[\mathbf{V}_\theta(\boldsymbol{\theta}, \boldsymbol{\varphi})]\} \quad , \tag{2.6}$$

where $\mathbf{V}_\theta$ is the parameter variance-covariance matrix, and ψ is a selected metric of $\mathbf{V}_\theta$ that represents the criterion for the experimental design. $\mathbf{V}_\theta$ is evaluated as:

$$\mathbf{V}_\theta(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \left\{ (\mathbf{V}_\theta^0)^{-1} + \sum_{k=1}^{N_{sp}} \sum_{i=1}^{N_y} \sum_{j=1}^{N_y} s_{i,j} \left[\frac{\partial \hat{y}_i(t_k)}{\partial \theta_l} \frac{\partial \hat{y}_j(t_k)}{\partial \theta_m} \right]_{l,m=1:1:N_\theta} \right\}^{-1} \quad , \tag{2.7}$$

where $s_{i,j}$ is the i,j^{th} element of the inverse of the measurement error covariance matrix, N_{sp} is the number of sampling intervals, N_y is the number of measured variables, N_θ is the number of parameters, and $\mathbf{V}_\theta^0$ is the preliminary parameter variance-covariance matrix.

To compare the magnitude of different parameter variance-covariance matrices, a selected metric ψ is evaluated through different criteria:

- A-Optimal: minimizes the trace of $\mathbf{V}_\theta(\boldsymbol{\theta}, \boldsymbol{\varphi})$, and thus minimises the dimensions of the enclosing box around the joint confidence region;
- E-Optimal: minimise the largest eigenvalue of $\mathbf{V}_\theta(\boldsymbol{\theta}, \boldsymbol{\varphi})$, and thus minimises the size of the major axis of the joint confidence region;
- D-Optimal: minimise the determinant of $\mathbf{V}_\theta(\boldsymbol{\theta}, \boldsymbol{\varphi})$, and thus minimises the volume of the joint confidence region.

A geometrical interpretation of each criterion is presented in Figure 2.1. One of the main reasons why MBDoE is used is to improve the precision of the parameter estimates. To

determine if a parameter has been estimated with sufficient precision, a t-test is used. The t-value for each parameter is calculated at $(1-\alpha)\%$ confidence level as:

$$t_i = \frac{\hat{\theta}_i}{95\%\ confidence\ interval_i} = \frac{\hat{\theta}_i}{t\left(1-\frac{\alpha}{2}, N-N_\theta\right)\sqrt{v_{ii}}} \quad , \tag{2.8}$$

where $\hat{\theta}_i$ is the parameter estimate, v_{ii} is the ii^{th} term of the variance-covariance matrix, N is the number of samples and N_θ the number of parameters. A statistically satisfactory parameter estimation is reached when a parameter t-value t_i is greater than the reference t-value t_{ref} with $(1-\alpha)\%$ confidence level and $(N-N_\theta)$ degrees of freedom (Franceschini and Macchietto, 2008):

$$t_{ref} = t(1-\alpha, N-N_\theta) \quad . \tag{2.9}$$

In this work, a D-Optimal criterion was used and the significance level $\alpha\%$ was set to *5%*.

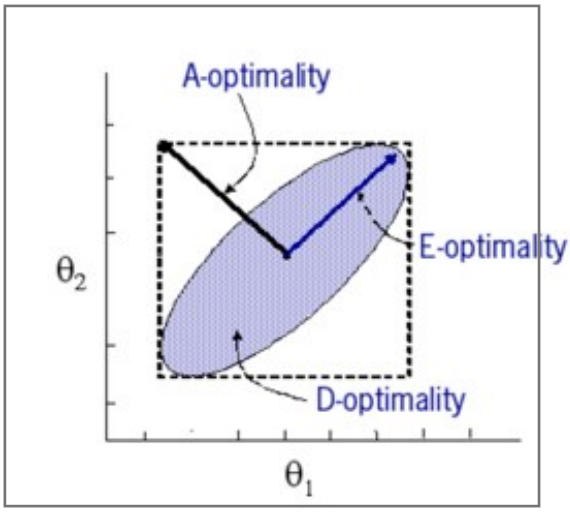


Figure 2.2. *Geometric interpretation of the various criterions*

MBDoE was performed on gPROMS®, within the `Experiment Design` entity, more specifically in the `…to be designed` section. There, the user can provide guesses for the initial conditions, the sensor and the values of manipulated variables that later are to be optimally determined by the MBDoE algorithm. Once the design is set, it is executed on the `Experiment Design` entity, and the results are obtained. After that, a new experiment with the MBDoE results is set in the `Experiments` entity, and its parameter estimation together with an evaluation of the precision of the parameter estimates is done in the `Model Validation` entity, where the parameters chosen to be estimated must be selected with their range of uncertainty, measured as an upper and lower value.

2.2. Selecting a subset of parameters to estimate

As in many practical situations model parameters are not all estimable, or obtaining sufficient data to estimate all the parameters reliably is too expensive or time consuming, it is often useful to determine which parameters can and should be estimated from the available experimental data. To do so, different simplified models (SMs), which have a smaller number of parameters, can be generated (Wu at al., 2011), each one having its own parameter subset that will be estimated from the experimental data, while the other parameters remain fixed at their nominal values.

Deciding which parameters to fix and which to estimate is referred as parameter subset selection. Several methods in the literature for selecting appropriate parameter subsets to estimate produce a ranked list of parameters, with the parameters appearing on the top being the ones that are the most important to be estimated so that good predictions can be obtained, while the parameters appearing on the bottom are the least important ones. However, a lot of these methods suffer from arbitrary cutoff values that are used to determine how many parameters to estimate and how many to fix (Thompson et al., 2009).

This problem motivated Wu et al., (2011) to propose a criterion based on mean squared error (MSE) to determine the number of parameters to be estimated from a ranked list. The mean squared error is the sum of the squared bias and variance, and it is a convenient measure for the quality of the model predictions. Using the definitions given in the original article, we will call *simplified models* (SMs), the models where only a subset of parameters is estimated from experimental data (and the rest is fixed) as compared to *extended models* (EMs), i.e. the original models where all parameters are estimated from available data. SMs can sometimes give better predictions, in terms of mean squared error, than EMs, especially if the available data for parameter estimation is noisy and limited.

When using a SM, some parameters are omitted from the estimation procedure, leading to an increase in bias in the parameter estimates, which sometimes can negatively affect model predictions. On the other hand, it may lead to a reduction in the variance of model predictions. Therefore, while using an SM may result in large bias in parameter estimates, overfitting or estimating too many parameters decreases the model predictive ability and may lead to large variances in model predictions, since when too many parameters are estimated using limited data, the high levels of uncertainty associated with the parameter estimates result in large variances in the model predictions (Thompson et al., 2009). A SM gives better predictions than the EM when the variance reduction outweighs the bias.

The orthogonalization method, which is used for ranking model parameters to most to least important, and the MSE-based method, used to select model parameters will be presented next.

2.3. Orthogonalization method

The orthogonalization method was first introduced by Yao et al. (2003) for selecting estimable parameters in a complex dynamic reactor model containing 50 parameters. In their approach, Yao et al. used the average values of response variables and initial parameter values for scaling. Later, Thompson et al. (2009) refined this algorithm by incorporating uncertainty-based scaling factors.

Referring to a model described by equations (1.3) and (1.4), the orthogonalization method relies on a sensitivity matrix **S** containing partial derivatives of the model predictions with respect to the model parameters:

$$\mathbf{S} = \begin{pmatrix} \left.\frac{\partial \hat{y}_{11}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{11}}{\partial \theta_p}\right|_{u_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{1n}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{1n}}{\partial \theta_1}\right|_{u_1} \\ \left.\frac{\partial \hat{y}_{21}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{21}}{\partial \theta_p}\right|_{u_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_p}\right|_{u_1} \\ \left.\frac{\partial \hat{y}_{11}}{\partial \theta_1}\right|_{u_2} & \cdots & \left.\frac{\partial \hat{y}_{11}}{\partial \theta_p}\right|_{u_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_1}\right|_{u_2} & \cdots & \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_p}\right|_{u_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_1}\right|_{u_r} & \cdots & \left.\frac{\partial \hat{y}_{dn}}{\partial \theta_p}\right|_{u_r} \end{pmatrix} , \qquad (2.10)$$

where the dimensions of **S** are $[N\ X\ p]$, with $N=dnr$, where d is the number of response variables, n is the number of measurement times, r is the number of experimental runs and p is the number of parameters. In this sensitivity matrix, each column is associated with a particular parameter and each row is associated with prediction of a particular measured value that will be used for parameter estimation.

For models that have analytical solutions, the derivatives in Equation (2.10) can be obtained directly through analytical methods. However, for more complex nonlinear ordinary differential equation (ODE) models, the elements of the sensitivity matrix **S** must be computed numerically. This can be done either by using finite difference approximations with perturbed parameter values (Saltelli et al., 2000) or by solving sensitivity equations (Leis and Kramer, 1988).

In order to apply the method, the sensitivity coefficients of **S** (i.e. elements of matrix **S**) should be properly scaled so that they are dimensionally consistent and can be meaningfully compared (Thompson et al., 2009). A new matrix **Z** is then formed, called scaled sensitivity matrix, on which each element of **S** is multiplied by scaling factors:

$$Z_{ilm,j} = \frac{\partial \hat{y}_{ilm}}{\partial \theta_j} \frac{s_{\theta j0}}{s_{yi}} \quad . \tag{2.11}$$

These scaled sensitivity coefficients are calculated for each of the $i = 1 \dots d$ response variables, each of the $l = 1 \dots n$ measurement times and each of the $m = 1 \dots r$ experimental runs. Scaling factors must be chosen carefully, because scaling can strongly affect the outcome of the parameters ranking. Scaling factor $s_{\theta j0}$ accounts for uncertainties in the initial guess for parameters and scaling factor s_{yi} reflects the uncertainty associated with the measured responses. Factor s_{yi} can be determined from replicate experiments or information about sensor suppliers, since it reflects the inaccuracies of different measurements (McLean et al., 2012), while factor $s_{\theta j0}$ reflects how far the modeler is willing to allow the particular parameter to move away from its initial guess.

The orthogonalization method, showed in Table 2.1, relies on the scaled sensitivity matrix to rank parameters from the most estimable to the least estimable. The rank of each model parameter is determined based on the influence of the parameter on model predictions and on correlations with other parameters.

Table 2.2. *Orthogonalization algorithm*

1.	Compute the magnitude (Euclidean norm) of each column in the **Z** matrix. Select the column with the largest magnitude as the most estimable parameter. Set $k = 1$.
2.	Put the *k* selected columns from **Z** that correspond to parameter that have been ranked in the matrix $\mathbf{X}_\mathrm{k}$.
3.	Use $\mathbf{X}_\mathrm{k}$ to predict columns in **Z** using ordinary least squares (OLS): $$\hat{\mathbf{Z}}_\mathrm{k} = \mathbf{X}_\mathrm{k}\left(\mathbf{X}_\mathrm{k}^\mathrm{T}\mathbf{X}_\mathrm{k}\right)^{-1}\mathbf{X}_\mathrm{k}^\mathrm{T}\mathbf{Z} \quad , \tag{2.12}$$ And calculate the residual matrix: $$\mathbf{R}_\mathrm{k} = \mathbf{Z} - \hat{\mathbf{Z}}_\mathrm{k} \quad , \tag{2.13}$$
4.	Calculate the magnitude of each column in $\mathbf{R}_\mathrm{k}$. The (k+1)th most estimable parameter corresponds to the column in $\mathbf{R}_\mathrm{k}$ with the largest magnitude.
5.	Increase the iteration counter *k* by one and repeat steps 2-4 until all parameters are ranked.

With all parameters ranked, parameters that stay on top of the ranked list are the ones with strong and independent influence on one or more model predictions, being then the most important parameters to be estimated, while on the bottom of the ranked list are the less

important parameters, on which can possibly give estimability problems. One advantage of scaling based on the initial uncertainties of parameters is that parameters with large uncertainties are more likely to be selected for estimation, while those that are known with high precision a-priori are placed on the bottom of the ranked list, and most of the times excluded from the estimation process.

The steps of the method will be discussed in the following paragraph.

2.3.1. Graphical visualization of the orthogonalization method

The ranking of the parameters starts by calculating the magnitude of each column of the scaled sensitivity matrix **Z**. In Figure 2.2, columns of matrix **Z** are visualized in form of plots, showing the scaled sensitivity coefficients (i.e. the elements of the matrix **Z**) in function of time. The following example is taken from a kinetic model that will be further discussed and introduced in Chapter 3. The model has four parameters ($\theta_1, \theta_2, \theta_3$ and θ_4) and two output variables (x_1 and x_2) that were measured every hour for the duration of 10 hours. For the sake of interpretability, the response variables x_1 and x_2 were shown in different plots.

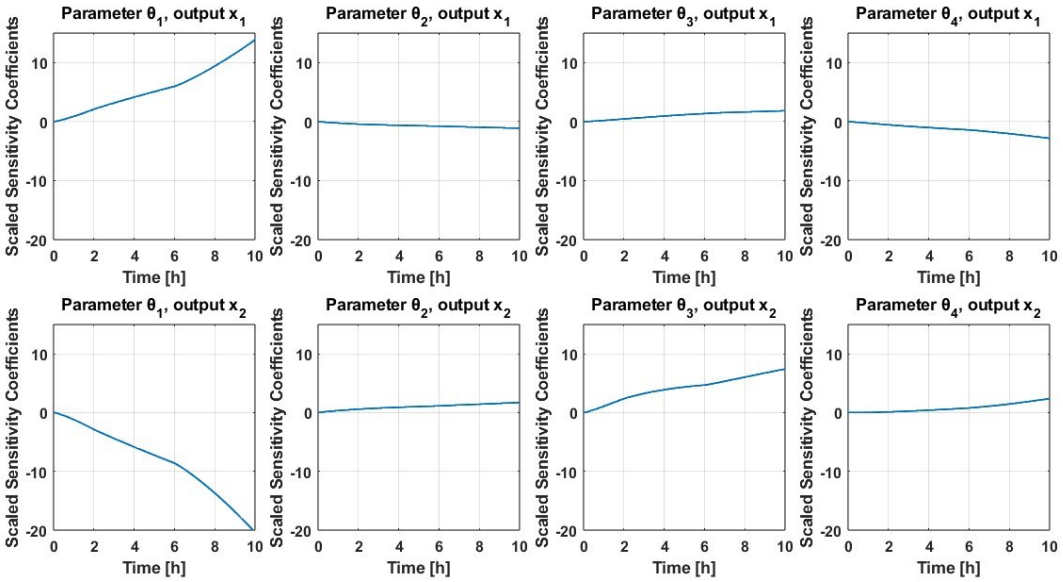


Figure 2.3. *Visualization of matrix* **Z***. The columns in this figure represent the columns of the matrix* **Z**

From Figure 2.2 it is easy to see graphically that column 1, which corresponds to parameter θ_1, presents the largest magnitude, therefore the method would select parameter θ_1 to be the most estimable (Step 1). This would lead to use the first column of matrix **Z** to predict the columns of **Z** using ordinary least squares (Step 2). Figure 2.3 shows the prediction of **Z** using the first column of **Z** (Step 3). As expected, the column of the most estimable parameter previously selected is perfectly predicted, since to predict it, the same exact column of **Z** was used.

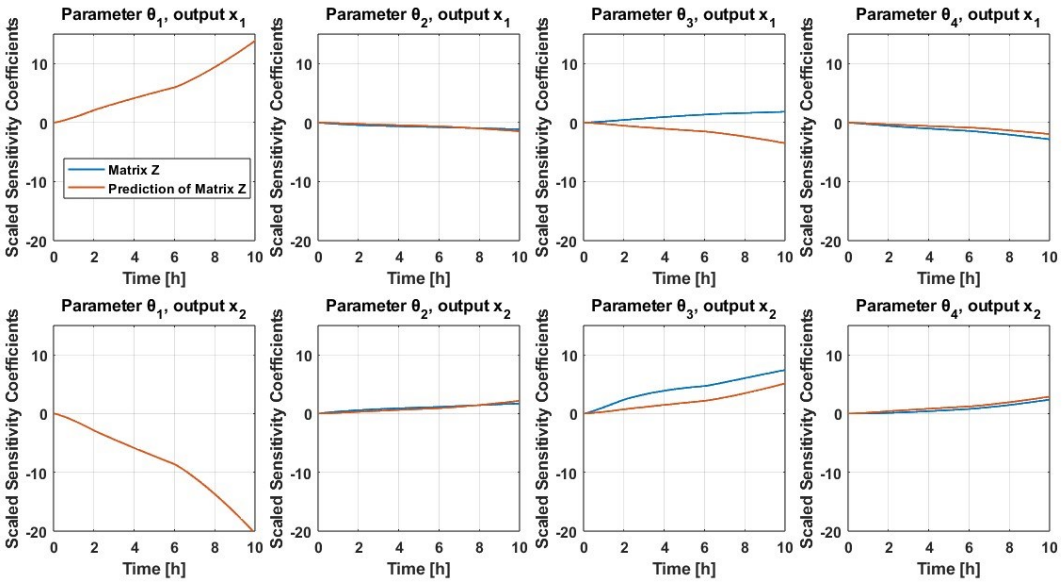


Figure 2.3. *Visualization of the predicted matrix using ordinary least squares*

The next most estimable parameter will be then the one with the largest magnitude among the other unranked parameter columns. It is already seen in Figure 2.3 that column 3 of matrix **Z** has the largest prediction residuals (i.e. the difference between the original column of matrix **Z** and the predicted column of **Z** is larger than the columns 2 and 4). Figure 2.4 shows the computation of the residual matrix $\mathbf{R}_k$ (Step 3).

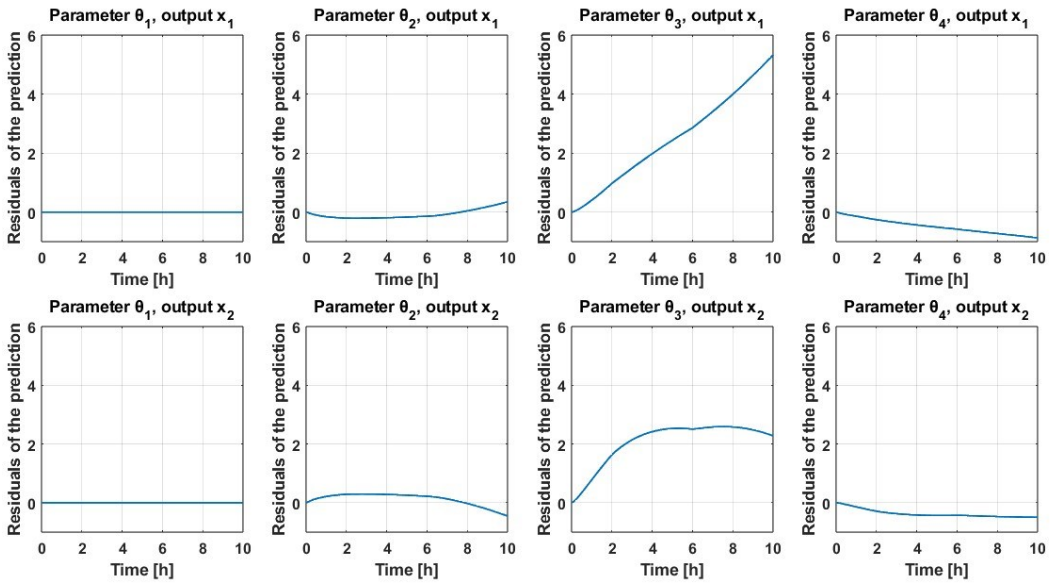


Figure 2.4. *Visualization of the residual matrix*

Confirming the visual results of Figure 2.3, Figure 2.4 shows that column 3 of matrix $\mathbf{R}_k$, which corresponds to parameter θ_3, has the largest magnitude (Step 4), therefore parameter θ_3 should be selected as the next most estimable parameter. This procedure should be repeated until all the parameters are ranked.

The orthogonalization method therefore calculates the magnitude of the columns of $\mathbf{Z}$ and $\mathbf{R}_k$ to assess the "net" influence of a parameter on the model responses, while the regression step (i.e. ordinary least squares prediction) is used to assess the correlation between parameters, being the columns of $\mathbf{Z}$ that are most similar to the predicted columns of $\mathbf{Z}$ the ones that are more correlated.

2.4. MSE-based parameter selection method

The MSE-based method, developed by Wu et al., (2011), determine the optimal number of parameters to estimate from the ranked list obtained via orthogonalization, so that the best predictions (i.e. with the lowest expected MSE) are obtained.

The method is developed using univariate linear models, but the extension to multivariate nonlinear models is quite straightforward, by assuming that the statistical properties of the nonlinear model can be adequately characterized by its linearised representation (Wu et al., 2011).

Consider an EM that can be described by the following univariate linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad , \tag{2.14}$$

where $\mathbf{Y} \in \mathbb{R}^n$ denotes a vector of measured responses, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is matrix of independent variable settings, $\mathbf{X}_1 \in \mathbb{R}^{n \times p_1}$, $\mathbf{X}_2 \in \mathbb{R}^{n \times (p-p_1)}$ are submatrices of $\mathbf{X}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of all unknown parameters, $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{(p-p_1)}$ are subvectors of $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of additive random noise terms, where n is the total number of measurements, p is the total number of parameters in the EM, p_1 is the number of parameters in the SM, and $(p - p_1)$ is the number of parameters excluded from the SM. Additionally, the following least-squares assumptions (Beck and Arnold, 1977) are made:

- $\mathbf{X}_1$ and $\mathbf{X}_2$ are perfectly known;
- Stochastic component $\boldsymbol{\varepsilon}$ is independently and identically distributed with mean zero and constant variance σ^2.

A particular SM is of the form:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e} \ , \tag{2.15}$$

where $\mathbf{e} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ is the stochastic component combined with any model mismatch. Note that the SM is nested within the EM. The MSE is defined as the expected squared difference between the model prediction, $\widehat{\mathbf{Y}}$, and the noise-free response of the process, $\mathbf{Y}_{\text{true}} = \mathbf{X}\boldsymbol{\beta}$ (Rice, 1995):

$$\begin{aligned} MSE\left(\widehat{\mathbf{Y}}\right) &= E\left(\left(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{true}}\right)^T\left(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{true}}\right)\right) \\ &= \left(E\left(\widehat{\mathbf{Y}}\right) - \mathbf{Y}_{\text{true}}^{\text{T}}\right)\left(E\left(\widehat{\mathbf{Y}}\right) - \mathbf{Y}_{\text{true}}\right) + tr\left(Cov\left(\widehat{\mathbf{Y}}\right)\right) \ , \end{aligned} \tag{2.16}$$

where $E(\cdot)$ denotes the expected value, $Cov(\cdot)$ denotes the variance-covariance matrix and $tr(\cdot)$ denotes the trace. The second line of Equation (2.16) shows that the MSE is equal to the squared bias $\left(E\left(\widehat{\mathbf{Y}}\right) - \mathbf{Y}_{\text{true}}^{\text{T}}\right)\left(E\left(\widehat{\mathbf{Y}}\right) - \mathbf{Y}_{\text{true}}\right)$ plus the total variance $tr\left(Cov\left(\widehat{\mathbf{Y}}\right)\right)$ of the model predictions (Rice, 1995).

Using ordinary least squares, model predictions at the design points used for parameter estimation are:

$$\widehat{\mathbf{Y}}_{\text{S}} = \mathbf{X}_1(\mathbf{X}_1^{\text{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\text{T}}\mathbf{Y} = P_1\mathbf{Y} \ , \tag{2.17}$$

$$\widehat{\mathbf{Y}}_{\text{E}} = \mathbf{X}(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}\mathbf{Y} = P\mathbf{Y} \ , \tag{2.18}$$

where $P_1 = \mathbf{X}_1(\mathbf{X}_1^{\text{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\text{T}}$ and $P = \mathbf{X}(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}$. Subscripts "S" and "E" indicate the use of a SM and EM, respectively. When unknown parameters in the EM (Equation 2.14) and SM (Equation 2.15) are estimated using ordinary least squares (OLS), the expected total MSE for predictions at the design points is (Beck and Arnold, 1977):

$$MSE_S = E\left(\left(\widehat{\mathbf{Y}}_{\text{S}} - \mathbf{X}\boldsymbol{\beta}\right)^T\left(\widehat{\mathbf{Y}}_{\text{S}} - \mathbf{X}\boldsymbol{\beta}\right)\right) = \sigma^2 p_1 + \boldsymbol{\beta}_2^{\text{T}}\mathbf{X}_2^{\text{T}}(\mathbf{I}_{\text{n}} - P_1)\mathbf{X}_2\boldsymbol{\beta}_2 \ , \tag{2.19}$$

$$MSE_E = E\left(\left(\widehat{\mathbf{Y}}_{\mathrm{E}} - \mathbf{X}\boldsymbol{\beta}\right)^T\left(\widehat{\mathbf{Y}}_{\mathrm{E}} - \mathbf{X}\boldsymbol{\beta}\right)\right) = \sigma^2 p \quad , \tag{2.20}$$

where $\mathbf{I}_{\mathrm{n}}$ is the $(n \times n)$ identity matrix. The first term on the right-hand side of Equation (2.19), $\sigma^2 p_1$, is the total variance in model predictions made at the design points. The second term, $\boldsymbol{\beta}_2^T \mathbf{X}_2^T (\mathbf{I}_{\mathrm{n}} - P_1)\mathbf{X}_2\boldsymbol{\beta}_2$, is the corresponding total bias (Wu et al., 2011).
The MSE-based method is used to determine whether the SM or the EM is expected to give predictions with lower MSE at the design points used for parameter estimation. This strategy relies on a critical ratio R_C, which was defined by Wu et al., (2007) as:

$$R_C = \frac{\boldsymbol{\beta}_2^{\mathrm{T}} \mathbf{X}_2^{\mathrm{T}} (\mathbf{I}_{\mathrm{n}} - P_1)\mathbf{X}_2\boldsymbol{\beta}_2}{(p - p_1)\sigma^2} \quad . \tag{2.21}$$

The critical ratio R_C is the total squared bias in the model prediction, introduced by removing parameters from the EM, divided by the variance reduction in the model predictions, due to few parameters in the SM considered. Therefore, the inequality:

$$R_C < 1 \quad , \tag{2.22}$$

is a necessary and sufficient condition for $MSE_S < MSE_E$, meaning that model predictions obtained using the SM have a smaller mean squared error than those from the EM. This then implies that the SM is preferable to the EM for making predictions (Wu et al., 2007).
Unfortunately, R_C cannot be calculated directly, because the true parameter values $\boldsymbol{\beta}_2$ and the true noise variance σ^2 are unknown. These values can then be replaced with the parameter estimates $\widehat{\boldsymbol{\beta}}_2$ and the noise variance estimate s_E^2 obtained using the EM and the available data, by fitting the EM using OLS, resulting in the following estimator for R_C:

$$r_C = \frac{\widehat{\boldsymbol{\beta}}_2^{\mathrm{T}} \mathbf{X}_2^{\mathrm{T}} (\mathbf{I}_{\mathrm{n}} - P_1)\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2}{(p - p_1)s_E^2} = \frac{(SSE_S - SSE_E) \,/\, (p - p_1)}{(SSE_E) \,/\, (n - p)} \quad , \tag{2.23}$$

where SSE is the sum of squared residuals and n the number of measurements. When σ^2 is assumed to be known, from prior information about the variability of the response, the estimator for R_C becomes:

$$r_C = \frac{\widehat{\boldsymbol{\beta}}_2^{\mathrm{T}} \mathbf{X}_2^{\mathrm{T}} (\mathbf{I}_{\mathrm{n}} - P_1)\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2}{(p - p_1)\sigma^2} = \frac{(SSE_S - SSE_E) \,/\, (p - p_1)}{\sigma^2} \quad . \tag{2.24}$$

The MSE-based method for parameter selection, however, is based on a corrected critical ratio R_{CC}, which is defined as:

$$R_{CC} = \frac{(MSE_S - MSE_E) \,/\, n}{\sigma^2} = \frac{p - p_1}{n}(R_C - 1) \quad , \tag{2.25}$$

since when comparing two SM_s with different parameter numbers, and thus different values of p_1, the SM with the lower value of R_C may not correspond to the lower MSE_S. This can be seen from the reduction in the total MSE at the design points, when a particular SM is used:

$$\Delta MSE = MSE_S - MSE_E = \sigma^2(p - p_1)(R_C - 1) \quad . \tag{2.26}$$

Because of the $(p - p_1)$ term in Equation (2.26), the SM with lower value of R_C may not be the one with the lower MSE_S (Wu et al., 2011). As a result, the corrected critical ratio R_{CC} is used for comparing several models with different number of parameters.

For the EM, the value of R_{CC} is zero, since for the EM, p is equal to p_1. In a case where the EM has the smallest MSE, the R_C for all the SM_s will be greater than 1, and their corresponding R_{CC} will be positive. In this situation, the EM will give better predictions in terms of MSE than the SM_s. In situations where some of the SM_s have a lower MSE than the EM, the R_C for these SM_s will be lower than 1, and their corresponding R_{CC} will be negative, indicating that these SM_s will give better predictions, in terms of MSE, than the EM. The SM with the lowest value of R_{CC} will be then the one that gives the best predictions.

Expression of r_C in Equations (2.23) and (2.24) can be used to calculate an estimate of R_{CC}. Unfortunately, these simple estimators for r_C can have a large bias and variance (Kubokawa et al., 1993). When σ^2 is unknown, the following truncated estimator should be used:

$$r_{C,Kub} = max\left(\frac{n - p - 2}{n - p}\, r_C - 1\,, \qquad \frac{2(n - p - 2)}{(p - p_1 + 2)(n - p)}\, r_C\right) \quad , \tag{2.27}$$

where the subscript *Kub* indicates that the estimator was derived using the improved estimator developed by Kubokawa et al., (1993). The r_C in Equation (2.27) is obtained from Equation (2.23). When σ^2 is known, the appropriate truncated estimator is:

$$r_{C,Kub} = max\left(r_C - 1, \frac{2}{p - p_1 + 2}\, r_C\right) \quad , \tag{2.28}$$

where r_C is obtained from Equation (2.24). The MSE estimate obtained from the truncated estimators in Equations (2.27) and (2.28) is lower than with the original r_C estimators in Equations (2.23) and (2.24). As a result, to select the best model that gives better prediction the following estimate for the corrected critical ratio R_{CC} is recommended:

$$r_{CC} = \frac{p - p_1}{n}\left(r_{C,Kub} - 1\right) \quad . \tag{2.29}$$

The candidate model with the lowest value of r_{CC} will be then the one that gives the best predictions in terms of MSE at the experimental points. Table 2.2 shows the algorithm to be implemented:

Table 3.2. *MSE-based algorithm to determine the optimal number of parameters to estimate*

1.	Rank model parameters from most estimable to least estimable using the orthogonalization method in Table 2.1.
2.	Use weighted least squares (WLS) regression to estimate the first parameter from the list, with all other parameters fixed at initial guesses. Next, estimate the top two parameters, followed by the top three parameters and so on, until all of the ranked parameters have been estimated. Denote the value of the objective function with the top k parameters estimated and the remaining $p - k$ parameters held fixed as J_k.
3.	Compute the critical ratio: $$r_{C,k} = \left(J_k - J_p\right)/(p - k) \ , \tag{2.30}$$ For $k = 1,2, \dots, p - 1$
4.	For each value of k, compute the corrected critical ratio: $$r_{CC,k} = \frac{p - k}{N}\left(r_{CKub,k} - 1\right) \ , \tag{2.31}$$ where $$r_{C,Kub,k} = max\left(r_C - 1, \frac{2}{p - p_1 + 2}\, r_C\right) \ , \tag{2.32}$$
5.	Select the value of k corresponding to the lowest value of $r_{CC,k}$ as the appropriate number of parameters to estimate.

Parameters in the k^{th} SM, as highlighted in Table 2.2 (Step 2), are estimated by minimizing the weighted sum of squared residuals, of a weighted least squares objective function J, defined as:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{d}\sum_{l=1}^{n}\sum_{m=1}^{r}\left(\frac{y_{ilm} - \hat{y}_{ilm}(\boldsymbol{\theta})}{s_{yi}}\right)^2 \ , \tag{2.33}$$

where s_{yi} is the uncertainty of the i^{th} measured response variable, which was used to scale the sensitivity matrix in Equation (2.11). Note that on Table 2.2 (Step 4), computation of $r_{CC,k}$ is done with n set to N, which is the total number of responses for all variables combined, allowing the method to be used for multivariate models.

It may seem that the expression of Equation (2.30) for calculating the critical ratio r_C is different from the one derived in Equation (2.24), but they are equivalent, considering that the uncertainty factor s_{yi} is a variance. This can be better seen if Equation (2.24) is rearranged:

$$r_C = \frac{1}{(p - p_1)}\left(\frac{SSE_S}{\sigma^2} - \frac{SSE_E}{\sigma^2}\right) = \frac{\left(J_{p_1} - J_p\right)}{(p - p_1)} \quad . \tag{2.34}$$

In cases where it is impossible to estimate all p parameters simultaneously due to numerical difficulties, J_p can be approximated by estimating a sufficiently large number of parameters, where estimation of additional parameters does not produce a noticeable reduction in the objective function value (McLean et al., 2012). After determining the number of parameters to be estimated, it is crucial to compare the model's predictions with the experimental data to assess potential lack of fit, which may suggest underlying structural issues with the model.

2.5. Implementation of the methods

The implementation starts by first calculating the elements of the sensitivity matrix **S**, which are partial derivatives of the model predictions with respect to the model parameters. In this study, elements of **S** are calculated numerically using finite difference approximations, such as forward difference approximation, as shown is Equation (2.35).

$$\frac{\partial y}{\partial \theta} \approx \frac{\Delta y}{\Delta \theta} = \frac{y(\theta + \Delta\theta) - y(\theta)}{\Delta \theta} \quad , \tag{2.35}$$

where the parameter perturbation $\Delta\theta$ is kept as low as possible, around $1 - 2\%$ of the initial parameter value θ. This calculation is performed for one parameter at-time, meaning that the perturbation is done for just one parameter, while the others are held fixed at their nominal values. In this way, the first perturbed parameter will then be the parameter that corresponds to the first column of the scaled matrix **S**, then when a second parameter is perturbed, model predictions are calculated, sensitivity coefficients (Equation 2.35) are also calculated, filling the second column of the matrix **S**, and so on for all the model parameters.

Once the matrix **S** is formed, each element of **S** is multiplied by the scaling factors seen in Equation (2.11). The scaling factor that reflects the uncertainty of parameters, $s_{\theta j0}$, is calculated in this study as the difference between the upper bound and the lower bound of parameter values, decided a-priori, while the scaling factor that reflects the uncertainty in the measurements is set to be the standard error of the measurements, σ. When all elements of the scaled matrix **S** are scaled, the algorithm showed in Table 2.1 is implemented, ending up with a ranked list of parameters.

Once the ranked list is obtained, the MSE-based method can be applied. In this study, synthetic "experimental" data is obtained by simulation with the "true" values of the parameters, assumed to be known a-priori, and later adding a Gaussian noise with a determined constant variance. Upper and lower bound of parameters are then set as the constraints for the solver of parameter

estimation, and the steps of Table 2.2 are followed. Finally, an optimal number of parameters to be selected is found, by selecting the smaller value of $r_{CC,k}$, and graphical inspection of the predictions with respect to the experimental data is done to verify the success of the methods.

2.6. Robustness test

One may not be confident in the values of the initial guesses of the parameters that are required for the application of the orthogonalization method. Different values of initial parameter guesses can result in different values for the scaled sensitivity coefficients in **Z**, and it can lead to a different parameter ranking and/or in different parameters selected for estimation.

To assess the robustness of the results of the method proposed by Wu et al., (2011), and accordingly, the results of the orthogonalization method, a Monte Carlo technique was used. In particular, 100 sets of initial parameter guesses were chosen randomly from a uniform distribution for each of the model parameters, with lower and upper parameter bounds set to $\theta_{j0} \pm$ a pre-determined percentage of the parameter nominal value, that represents parameter uncertainty.

Then, 100 ranked lists were obtained by repeating the orthogonalization method, and the frequency of the rank for each parameter is reported graphically. The MSE-based method is then applied to the 100 ranked lists, each one representing a random set of initial parameter guesses, to determine the optimal number of parameters to estimate. The frequencies of the optimal number of parameters selected for estimation are also shown graphically.

This type of analysis can be valuable since it verifies if the methods agree for the most part on which parameters and how many of them are selected for estimation, based on the degree of uncertainty of parameters, and it also verifies if the first initial parameter guesses given by the modeler results in the most frequent outcome inside the specified range of parameter uncertainty. A drawback of this analysis is the computational time, since the orthogonalization and MSE-based methods need to be repeated for 100 times.

Chapter 3

Application of the methodology on a fermentation process model

This chapter firstly focuses on the application of the estimability method used to rank and select model parameters on a fermentation process model, considering different levels of measurement noise and uncertainty on parameter values. Then, the robustness of results is assessed using a Monte Carlo approach and the precision of the parameter estimates is further analysed. Finally, results are compared to other analysis such as LSA and GSA.

3.1. Fermentation process model

The methodology discussed in the previous chapter is applied to a biomass fermentation process for baker's yeast growth.

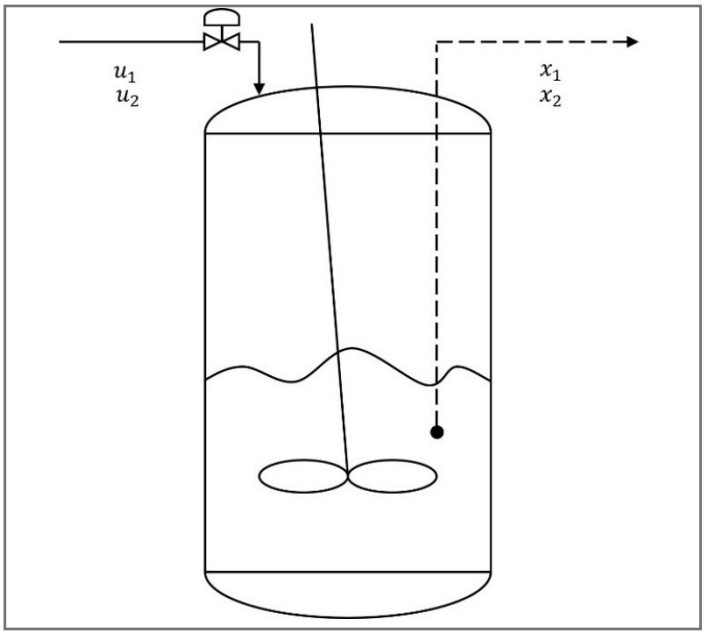


Figure 3.5. *Process scheme of a baker's yeast fermentation process*

Assuming Monod-type kinetics for biomass growth and substrate consumption, the system is described by the following set of DAEs presented in Galvanin et al. (2007):

$$\begin{cases} \dfrac{dx_1}{dt} = (r - u_1 - \theta_4)x_1 \\ \dfrac{dx_2}{dt} = -\dfrac{rx_1}{\theta_3} + u_1(u_2 - x_2) \\ r = \dfrac{\theta_1 x_2}{\theta_2 + x_2} \end{cases}, \tag{3.1}$$

where x_1 is the biomass concentration [g/L], x_2 is the substrate concentration [g/L], u_1 is the dilution factor [h^{-1}], and u_2 is the substrate concentration in the feed [g/L]. Variables u_1 and u_2 are manipulated variables, while x_1 and x_2 are measured outputs. The experimental conditions that characterize a particular experiment are the initial biomass concentration x_1^0 (range 1 – 10 g/L), the dilution factor u_1 (range 0.05 – 0.20 h^{-1}), and the substrate concentration in the feed u_2 (range 5 – 35 g/L). The initial substrate concentration x_2^0 is set to 0 g/L and cannot be manipulated for experimental design purposes. Both x_1 and x_2 can be measured during the experiment, therefore $\mathbf{x}(t) = \hat{\mathbf{y}}(t)$. The total duration of a single experiment is set equal to 10 h. It is assumed that the experimental run involves ten sampling times (i.e. times when the measured variables x_1 and x_2 were measured), one every hour. The true process conditions are presented in Table 3.1.

Table 3.4. *Process settings at nominal conditions*

Variable/Parameter	*Units*	*Nominal value*
t	h	10.00
x_1^0	g/L	5.00
x_2^0	g/L	0.00
θ_1	-	0.310
θ_2	-	0.180
θ_3	-	0.550
θ_4	-	5.000×10^{-2}

It is assumed that inputs $\mathbf{u}(t)$ can be manipulated and represented as a piecewise-constant profile over the switching intervals presented in Table 3.2.

Table 3.5. *Manipulated variables switching intervals and values*

	$\boldsymbol{u_1}\ [\boldsymbol{h^{-1}}]$	$\boldsymbol{u_2}\ [\boldsymbol{g/L}]$
$0 \leq t < 2\ h$	0.100	20.00
$2 \leq t < 6\ h$	0.200	10.00
$6 \leq t \leq 10\ h$	0.150	30.00

Once operating conditions are fixed, the methodology is then applied to the biomass fermentation process, with different levels of experimental data noise and uncertainty on the actual values of the parameters.

3.2. Methodology results with low experimental data noise

First, the estimability analysis is applied assuming that the available experimental data are affected by low level of noise. It is assumed that a single measurement sensor is present in the process, and that it measures both the response variables x_1 and x_2 with a constant variance of $\sigma^2 = 2.000\times10^{-2}$.

Therefore, the scaling coefficient that represent the uncertainty of the measures is set to be the standard error of the measurements, so $s_{yi} = \sigma = 0.141$ g/L. Results are then subdivided based on the level of uncertainty on parameter values.

3.2.1. Low uncertainty on parameter values

It is assumed that a low uncertainty on parameter values corresponds to when parameters stay in the range of $\pm$30.00% of the "true" parameter values, which are already known. The scaling factor reflecting parameter uncertainty is calculated as:

$$s_{\theta j0} = ub_{\theta_j} - lb_{\theta_j} \quad , \tag{3.2}$$

where ub_{θ_j} and lb_{θ_j} stand for upper bound of parameter θ_j and lower bound of parameter θ_j, respectively, and in the case of low uncertainty on parameter values, $lb_{\theta j} = \theta_j - (0.300 \times \theta_j)$ and $ub_{\theta j} = \theta_j + (0.300 \times \theta_j)$.

The initial guess of parameter values, along with their scaling factors of uncertainty is presented in Table 3.3.

Table 6.3. *"True" parameter values, initial parameter guesses and uncertainty factors*

Parameter	***"True" value***	***Initial guess***	***Uncertainty $s_{\theta j0}$***
θ_1	0.310	0.240	0.144
θ_2	0.180	0.220	0.132
θ_3	0.550	0.650	0.450
θ_4	5.000×10^{-2}	3.900×10^{-2}	2.340×10^{-2}

The orthogonalization method is then applied to the initial parameter guesses, following the procedure of Table 2.1. A ranked list of the parameters is obtained, as shown in Table 3.4.

Table 3.4. *Ranking obtained using the orthogonalization algorithm*

Parameter	***Rank***
θ_1	1
θ_2	3
θ_3	2
θ_4	4

The MSE-based method is then applied, and as expected, estimating additional parameters resulted in an improved fit of the data, as shown by the trends in the objective function J plotted in Figure 3.2.

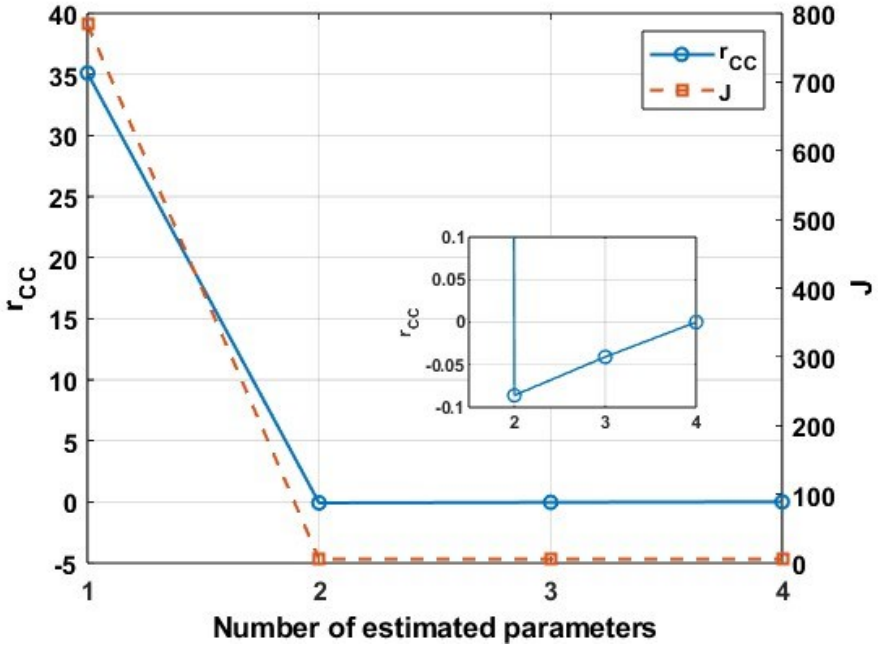


Figure 3.6: *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Note the different scales used on the vertical axis for $r_{CC,k}$ and the presence of a zoomed plot referring to the values of $r_{CC,k}$ close to zero. Case of low noise and low uncertainty on parameter values.*

An improvement in the objective function J is observed for up to two parameters estimated. When additional parameters are estimated, there is negligible improvement in the fit. The smallest value of $r_{CC,k}$ is when $k = 2$, clearly seen in the zoomed plot of Figure 3.2. Thus, estimating the first two parameters from the ranked list should give the best predictions with the lowest MSE. That is, parameters θ_1 and θ_3 should be estimated, and the other parameters should remain at their initial values specified in Table 3.3. Final results of the estimability analysis are summarized in Table 3.5.

Table 3.5. *Estimability analysis results. Case of low experimental noise and low uncertainty on parameter values.*

Parameter	***Rank***	***Selected***
θ_1	1	yes
θ_2	3	no
θ_3	2	yes
θ_4	4	no

Comparison between model predictions and experimental data, resulting from the estimation of the ranked list parameters is shown in Figure 3.3, along with the coefficient of determination R^2 in Table 3.6.

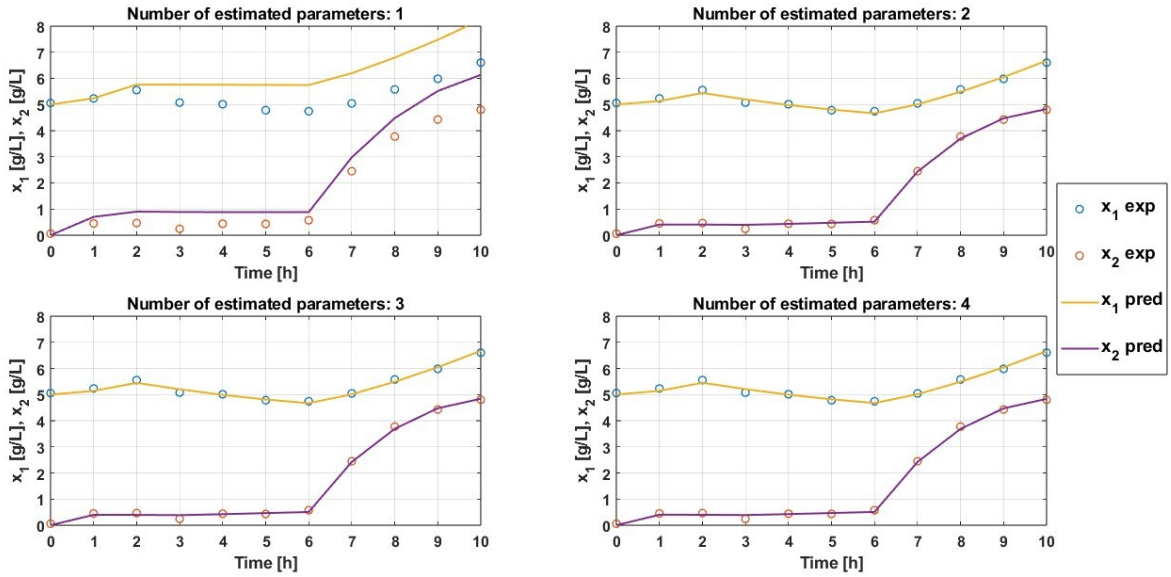


Figure 7.3. *Model predictions vs estimated parameters. "exp" refers to the experimental data and "pred" to the model predictions.*

Table 3.6. *R^2 values for each model output. Case of low experimental noise and low uncertainty on parameter values.*

Parameters estimated	x_1	x_2
1	-2.417	0.856
2	0.978	0.998

It can be seen from Figure 3.3 that the model predictions match the experimental data very well already when the first two ranked parameters are estimated, and that no significant improvements are made when more than two parameters are estimated. It can also be seen from

Table 3.6 that the R^2 of the two measured variables is very close to 1 when two parameters are estimated, indicating that the model fits the data very well when the parameters selected by the method are estimated.

3.2.2. High uncertainty on parameter values

When the uncertainty on parameter values is high, it was assumed that the initial parameter guesses can be taken from a range of $\pm 100.00\%$ of the parameters "true" value. Thus, the scaling factor reflecting the uncertainty is calculated by Equation (3.2), with $lb_{\theta j} = 0.00$ and $ub_{\theta j} = 2.000 \times \theta_j$. The initial guess of parameter values, along with their scaling factors of uncertainty is presented in Table 3.7.

Table 3.7. *"True" parameter values, initial parameter guesses and uncertainty factors.*

Parameter	***"True" value***	***Initial guess***	***Uncertainty $s_{\theta j0}$***
θ_1	0.310	0.120	0.240
θ_2	0.180	0.320	0.640
θ_3	0.550	0.800	1.600
θ_4	5.000×10^{-2}	8.000×10^{-2}	0.160

The results of the orthogonalization method are shown in Table 3.8.

Table 3.8. *Ranking obtained using the orthogonalization algorithm*

Parameter	***Rank***
θ_1	1
θ_2	3
θ_3	2
θ_4	4

The parameters ranking did not change from the one obtained with low uncertainty on parameter values. Trends of the objective function J and $r_{CC,k}$ values are shown in Figure 3.4.

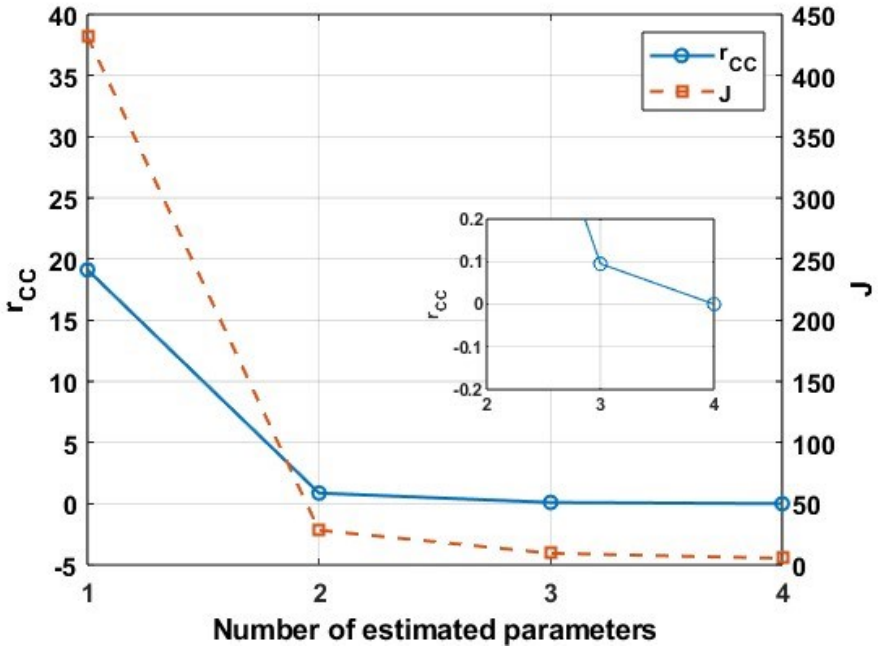


Figure 3.8. *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Case of low noise and high uncertainty on parameter values.*

Similarly to when the uncertainty on parameter values is low, negligible improvements of the objective function J are observed after estimating the first two ranked parameters. However, the lowest value of $r_{CC,k}$ is obtained when $k = 4$, thus according to the MSE-based method, all the model parameters should be estimated in order to obtain the best predictions with the lowest MSE. Final results are summarized in Table 3.9.

Table 3.9. *Estimability analysis results. Case of low experimental noise and high uncertainty on parameter values.*

Parameter	***Rank***	***Selected***
θ_1	1	yes
θ_2	3	yes
θ_3	2	yes
θ_4	4	yes

The model predictions with the corresponding number of estimated parameters are shown in Figure 3.5, and values of R^2 in Table 3.10.

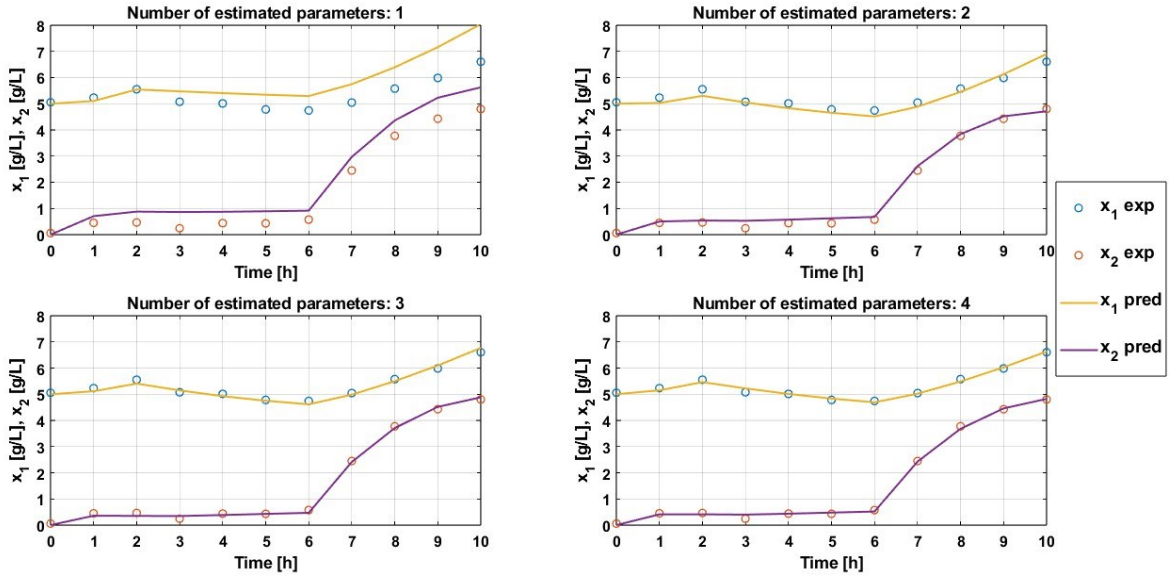


Figure 3.9. *Model predictions vs estimated parameters. Case of low noise and high parameter uncertainty.*

Table 3.10. *R^2 values for each model output. Case of low experimental noise and high uncertainty on parameter values.*

Parameters estimated	x_1	x_2
1	-0.766	0.910
2	0.886	0.994
3	0.962	0.998
4	0.981	0.998

When the uncertainty on parameter values is high, the MSE-based method tends to select all four parameters to be estimated, even though graphically the predictions seem to be quite good after estimating only the first two ranked parameters, with acceptable R^2 values. The methodology should be repeated after each estimation as the number of model parameters to be estimated is strongly influenced by the initial guess values.

3.3. Methodology results with high experimental data noise

The estimability analysis is applied for a case where experimental noise is high. It was assumed that the sensor that measures variables x_1 and x_2 has a constant variance of $\sigma^2 = 0.200$. Thus, the scaling factor that reflects the uncertainty of the measures is set to be $s_{yi} = \sigma = 0.447$ g/L.

3.3.1. Low uncertainty on parameter values

The same values of $s_{\theta j0}$ used in the previous case (§3.2.1) are also adopted in the case of high experimental noise, along with the initial parameter guesses, thus Table 3.3 serve as a reference for the case of high experimental noise too.
Results of the orthogonalization method are shown in the following ranked list:

Table 3.11. *Ranking obtained using the orthogonalization algorithm*

Parameter	***Rank***
θ_1	1
θ_2	3
θ_3	2
θ_4	4

Trends of the objective function J and $r_{CC,k}$ values are shown in Figure 3.6.

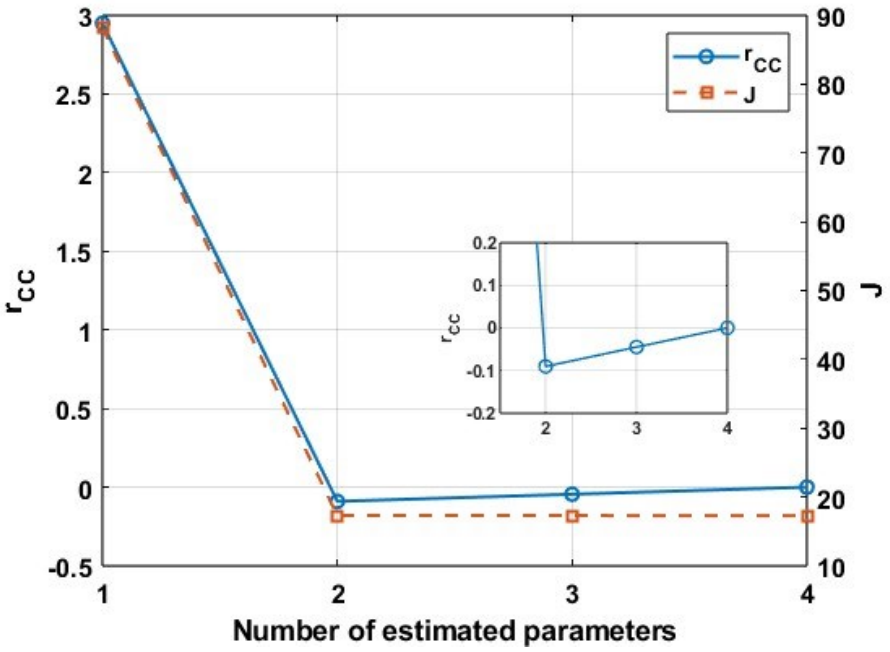


Figure 3.10. *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Case of high noise and low uncertainty on parameter values.*

The trend in the objective function J shows that there is no significant improvement in the fit of experimental data when more than two parameters are estimated. Moreover, the lowest value of $r_{CC,k}$ is when $k = 2$, so when the top two ranked parameters are estimated, while the others

are fixed at their initial guess, the model will give the best predictions in terms of MSE. The estimability results are summarized in Table 3.12, along with the model predictions in Figure 3.7 and R^2 values in Table 3.13.

Table 3.12. *Estimability analysis results. Case of high experimental noise and low uncertainty on parameter values.*

Parameter	***Rank***	***Selected***
θ_1	1	yes
θ_2	3	no
θ_3	2	yes
θ_4	4	no

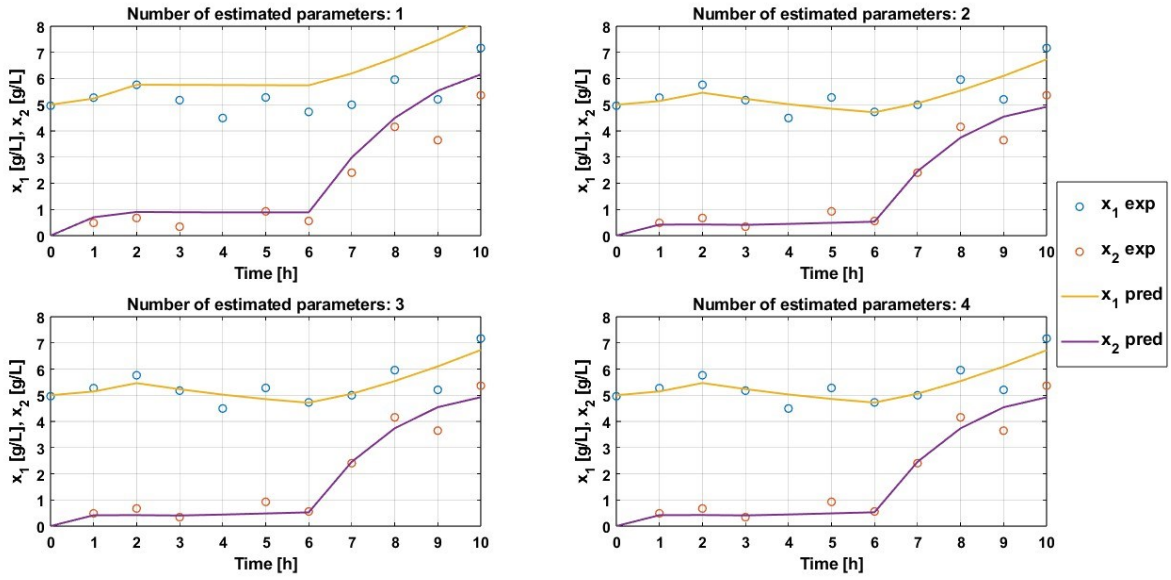


Figure 3.11. *Model predictions vs estimated parameters. Case of high noise and low uncertainty on parameter values.*

Table 3.13. *R^2 values for each model output. Case of high experimental noise and low uncertainty on parameter values.*

Parameters estimated	x_1	x_2
1	-1.176	0.831
2	0.670	0.952

Graphically, model predictions seem to not change significantly after estimating the first two ranked parameters, which is confirmed by checking the R^2 values, since they do not change significantly when two or more parameters are estimated. While the R^2 value regarding the

model output x_2 is very high, the R^2 value of x_1 is not very high even when all parameters are estimated, indicating that the model are not very accurate for the measured variable x_1, but still acceptable.

3.3.2. High uncertainty on parameter values

The same values of $s_{\theta j0}$ used in the previous case of low experimental noise (§3.2.2) are also adopted in the case of high experimental noise, along with the initial guesses of the parameters, referred in Table 3.7.

The ranking obtained from the orthogonalization method is showed in Table 3.14.

Table 3.14. *Ranking obtained using the orthogonalization algorithm.*

Parameter	***Rank***
θ_1	1
θ_2	3
θ_3	2
θ_4	4

And trends of the objective function J and $r_{CC,k}$ values are shown in Figure 3.8.

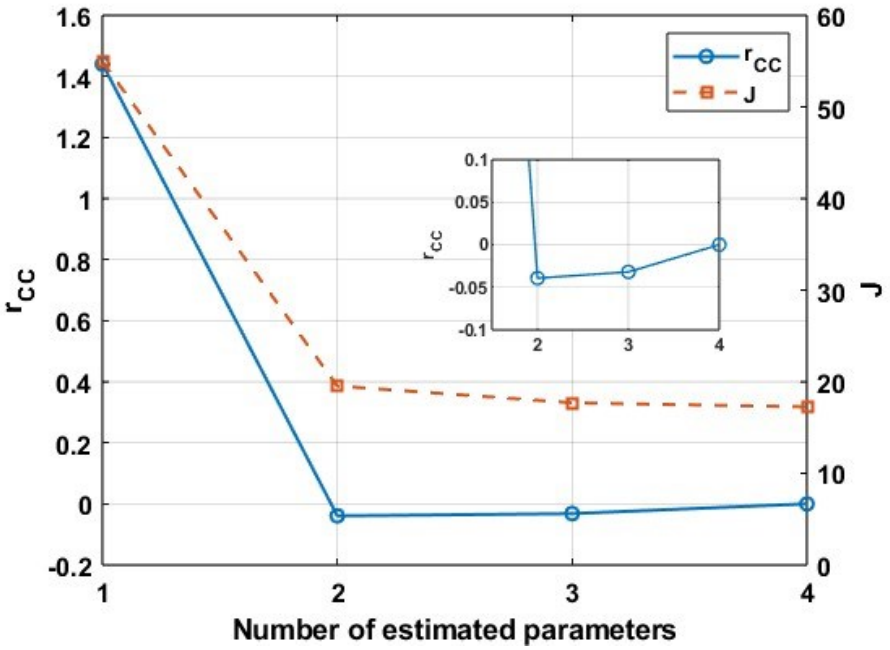


Figure 3.12. *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Case of high noise and high uncertainty on parameter values.*

Like the other cases, the trend on the objective function J tends to stabilize when more than two parameters are estimated, and the lowest value of $r_{CC,k}$ is when $k = 2$. The estimability results are summarized in Table 3.15, along with the model predictions in Figure 3.9 and R^2 values in Table 3.16.

Table 3.15. *Estimability analysis results. Case of high experimental noise and high uncertainty on parameter values.*

Parameter	***Rank***	***Selected***
θ_1	1	yes
θ_2	3	no
θ_3	2	yes
θ_4	4	no

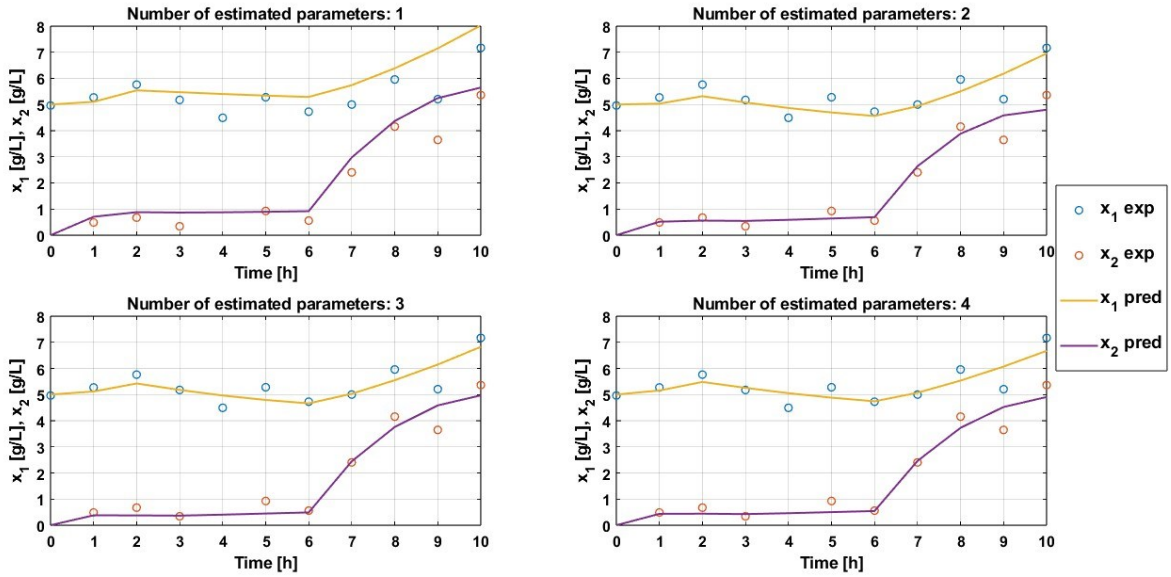


Figure 3.13. *Model predictions vs estimated parameters. Case of high noise and high uncertainty on parameter values.*

Table 3.16. *R^2 values for each model output. Case of high experimental noise and high uncertainty on parameter values.*

Parameters estimated	x_1	x_2
1	-0.238	0.877
2	0.627	0.946

where model predictions seem to not change significantly when more than two parameters are estimated, confirmed by the R^2 values.

3.4. Discussion

The methodology proposed by Wu et al., (2011) was successfully applied to a biomass fermentation process, where different levels of experimental data noise and uncertainty on the actual values of the parameters were analysed. Table 3.17 summarizes the results obtained.

Table 3.17. *Estimability analysis results for all cases*

Parameter	***Low noise***				***High noise***			
	Low uncertainty on parameter values		High uncertainty on parameter values		Low uncertainty on parameter values		High uncertainty on parameter values	
	Rank	Selected	Rank	Selected	Rank	Selected	Rank	Selected
θ_1	1	yes	1	yes	1	yes	1	yes
θ_2	3	no	3	yes	3	no	3	no
θ_3	2	yes	2	yes	2	yes	2	yes
θ_4	4	no	4	yes	4	no	4	no

Results show that the parameters ranking did not change for all the cases analysed, thus solidifying that the parameters ranking of importance is well determined. In the case of high noise in the experimental data, the MSE-based method tends to select two parameters for estimation. Trends of the objective function J for the case of high noise tend to stabilize at higher values than the cases of low noise, indicating than the data fit is worse when the noise in the experimental data is higher. Values of $r_{CC,k}$ on the other hand, for the cases of high experimental noise, tend to assume lower values when only one parameter is selected for parameter estimation, compared to the cases of low noise, indicating that the MSE-based method tends to prefer the estimation of less parameters when the experimental data noise is higher.

When the experimental data noise is low, the MSE-based method tends to select all the model parameters to estimate when the uncertainty on parameter values is high, indicating that the method tends to select more parameters for estimation from the ranked list when the experimental noise is low. Meanwhile, when the noise and the uncertainty on parameter values is low, the method tends to select the first two ranked parameters, therefore uncertainty on parameter values seems to have an important impact on the number of selected parameters, since when the modeller is very certain in the precision of the initial guess of the parameters, the method then tends to select less parameters for estimation.

3.5. Assessment of the methodology robustness

As presented in §2.6, a Monte Carlo technique is used to assess the robustness of the estimability analysis proposed by Wu et al., (2011). A number of 100 different initial guesses of the parameters from a random distribution inside a pre-defined range is considered, which is given by the initial uncertainty on parameter values that was assumed. Therefore, in cases where the uncertainty is high, the different sets of model parameters are taken from the interval of $\pm 100.00\%$ of the "true" parameter values, while if the uncertainty is low, parameter sets are taken from the interval of $\pm 30.00\%$ of the "true" parameter values.

3.5.1. Results with low experimental data noise

The frequency of the parameters ranking for the different levels of uncertainty on parameter values is assessed and shown in Figure 3.10.

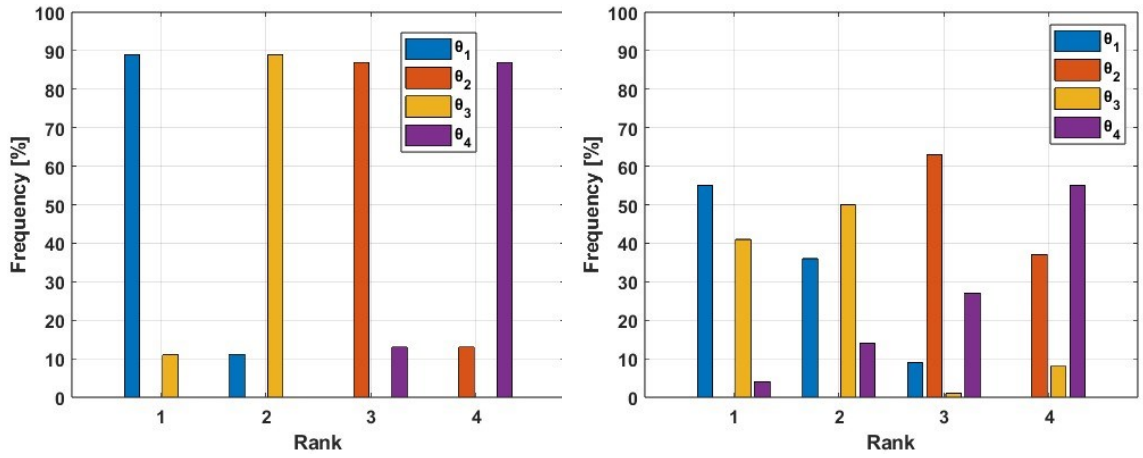


Figure 3.10. *Frequency of parameter rankings obtained using the algorithm in Table 2.1 with 100 random initial parameter guesses. Case of low uncertainty (a), and high uncertainty (b) on the parameter values.*

when the uncertainty on parameter values is low, the orthogonalization method agrees with a very high frequency on the position of the parameters in the ranked list (see Figure 3.10a). Conversely, when the uncertainty on parameters is high, the parameters ranking, depending on the initial guesses, is not very established (see Figure 3.10b). The robustness test highlights that parameters θ_1 and θ_3 are the most influential for the model predictions, staying most of the time in the first two positions, while parameters θ_2 and θ_4 are the less influential. Either way, for both cases the parameters ranking tends to agree to the one assessed in §3.2.

The frequency of the number of parameters selected for estimation is shown in Figure 3.11.

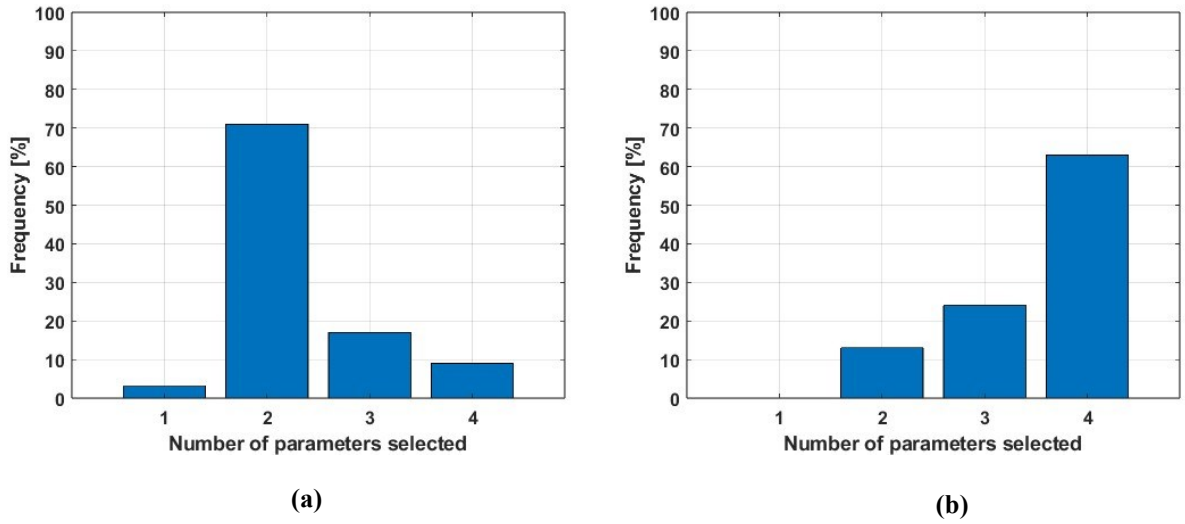


Figure 3.11. *Frequency of the number of parameters selected obtained using the algorithm in Table 2.2 with 100 random initial parameter guesses. Case of low uncertainty (a) and high uncertainty (b) on the parameter values.*

when the uncertainty on parameter values is low, the method selects the two top ranked parameters as the optimal number of parameters selected for estimation for roughly 70% of initial parameter guesses, while when the uncertainty on parameter values is high, for roughly 65% of the initial parameters guesses, the method selects four parameters to be estimated, confirming that what was found as a result in §3.2 is the most common outcome.

3.5.2. Results with high experimental data noise

The frequency of parameter rankings for both cases of low and high parameter uncertainty is shown in Figure 3.12.

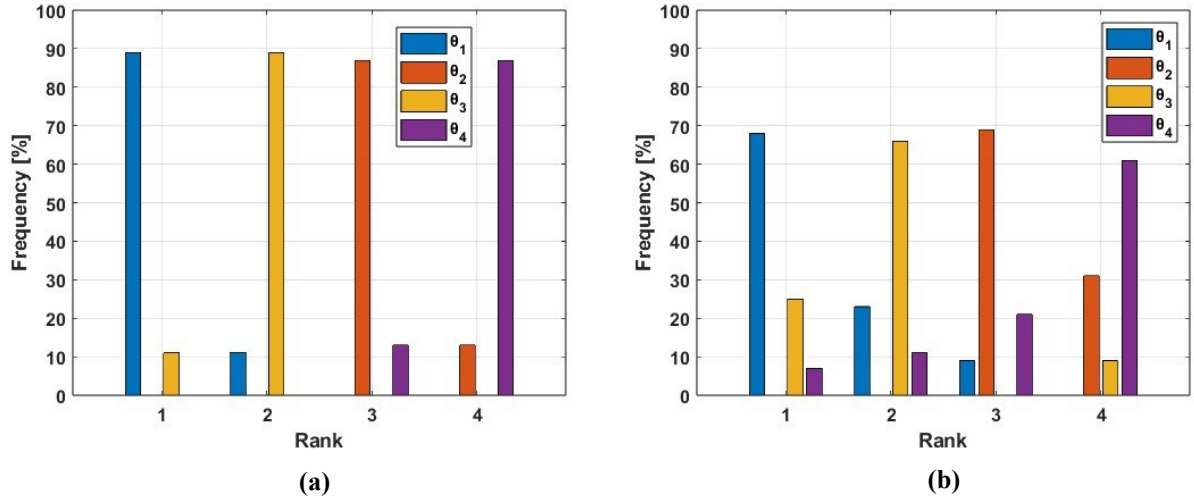


Figure 3.12. *Frequency of parameters rankings obtained using the algorithm in Table 2.1 with 100 random initial parameter guesses. Case of low uncertainty (a), and high uncertainty (b) on parameter values.*

As the case of low experimental noise, when the uncertainty on parameter values is low, all the parameters retain their original ranks almost 90% of the time, and when the uncertainty on parameter values is high, the ranks of the parameters are more variable, but for most of the time (almost 70% of the time), all the parameters retain their original ranks, as encountered in §3.3. The frequency of the number of parameters selected for estimation is shown in Figure 3.13.

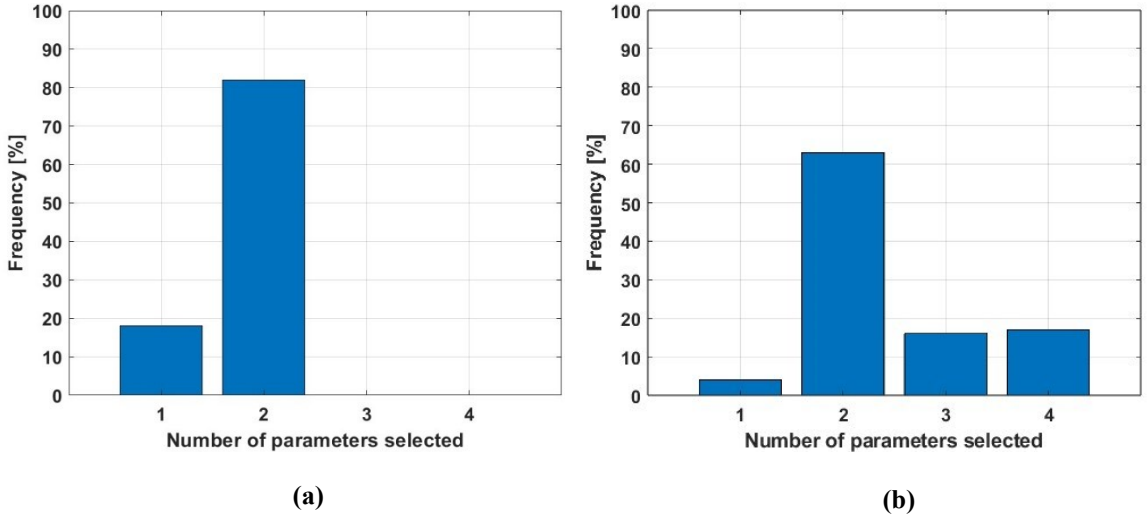


Figure 3.13. *Frequency of the number of parameters selected obtained using the algorithm in Table 2.2 with 100 random initial parameter guesses. Case of low uncertainty (a) and high uncertainty (b) on parameter values.*

In both cases the method selects most of the time two parameters as the optimal number of parameters to estimate, agreeing to the results obtained in §3.3.

3.5.3. Comments on results

The results of the robustness test indicate that the parameter rankings remain relatively stable across varying levels of experimental noise. However, the rankings are more significantly affected by changes in the levels of uncertainty associated with the parameter values. In particular, when the uncertainty on parameter values is low, the method tends to select most of the time the same rank for the parameters, while when the uncertainty on parameter values is high, the parameters ranking tend to vary more, even though it is clear that parameters θ_1 and θ_3 are the most influential, occupying almost every time the first two positions, while parameters θ_2 and θ_4 are the less important ones, almost always staying in the last two positions. Moreover, the test shows that when the uncertainty on parameter values is low, the MSE-based method tends to select less parameters for estimation (in this model of interest, the first two parameters) regardless of the experimental noise, while if the uncertainty on parameter values

is high, the method tends to select more parameters when the experimental noise is low (all the parameters) and less parameters when the experimental noise is high (the first two parameters). These results suggest that when values of parameter guesses are not close to the "true" values, the method tends to select more parameters to be estimated; in other words, the number of model parameters to be estimated depends on the initial values of parameter guesses.

3.6. Parameter precision

Results of the estimability analysis were analysed for sets of parameters that present an unusual ranking, with the scope of verifying if the parameter estimates obtained are statistically reliable, and therefore precise. If parameter estimates are not reliable, a second iteration of the methodology is applied using as initial guesses the parameter estimates obtained in the first iteration, and the precision of the second parameter estimates is assessed. If the estimates are still not precise, an MBDoE is then performed in order to improve the precision of the parameter estimates.

Parameter sets are taken from the robustness test when the experimental noise and the uncertainty on parameter values is high. The first parameter set is showed in Table 3.18, along with the results of the estimability analysis.

Table 3.18. *Estimability analysis results of the first anomalous parameter set.*

Parameter	***Initial guess***	***Rank***	***Selected***
θ_1	0.431	2	yes
θ_2	0.114	3	no
θ_3	0.641	1	yes
θ_4	5.400×10^{-3}	4	no

It can be seen an inversion on the parameter rankings between parameter θ_1 and θ_3, where usually parameter θ_1 retains the first position, but in this case, it assumes the second position of the ranking. Model predictions and R^2 values for different subsets of parameters to estimate are shown in Figure 3.14 and Table 3.19.

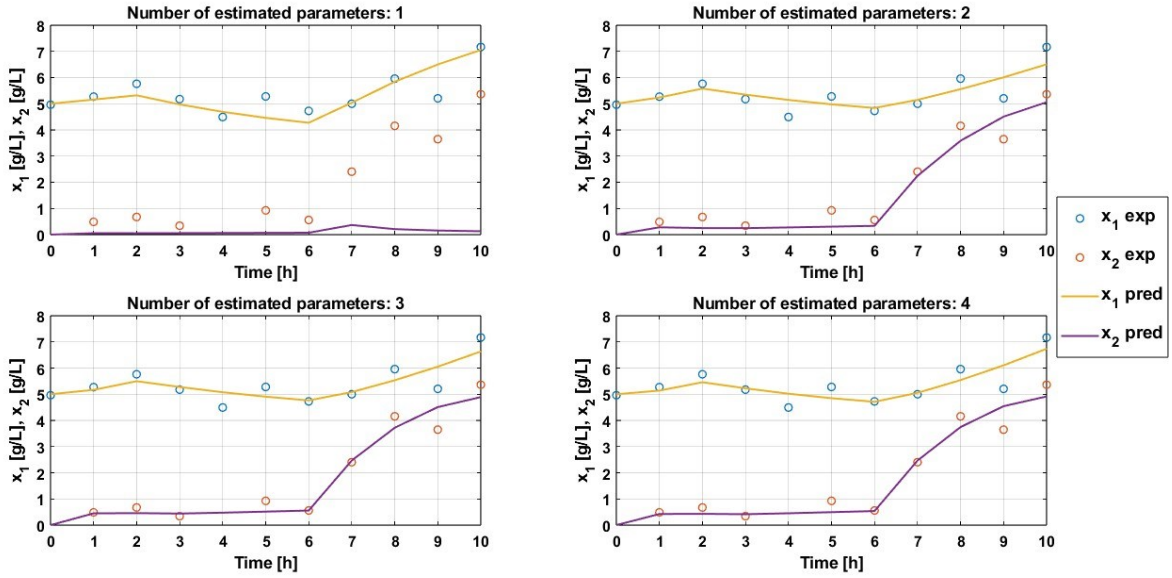


Figure 3.14. *Model predictions vs estimated parameters.*

Table 3.19. *R^2 values for each model output.*

Parameters estimated	x_1	x_2
1	0.459	-0.684
2	0.648	0.945

Considering what the MSE-based method suggests, only the first two parameters from the ranked list should be estimated. R^2 values also stabilize after estimating two parameters. Table 3.20 shows the parameter estimates of the first two ranked parameters, along with their corresponding t-values.

Table 3.20. *Estimation of model parameters chosen by the MSE-based method.*

Parameter	***Initial guess***	***Estimate***	**95% *CI***	**95% *t-value***
θ_1	0.431	0.240	7.800×10^{-3}	30.72
θ_3	0.641	0.455	2.840×10^{-2}	16.03
				$t_{ref} = 1.724$

As shown in Table 3.20, the parameter estimates t-values are higher than the reference one, indicating that the parameters are estimated with sufficient precision.

Another parameter set that has an anomalous ranking is the one presented in Table 3.21.

Table 3.21. *Estimability analysis results of the second anomalous parameter set.*

Parameter	***Initial guess***	***Rank***	***Selected***
θ_1	2.080×10^{-2}	2	yes
θ_2	0.232	4	no
θ_3	1.087	3	yes
θ_4	3.310×10^{-2}	1	yes

where parameter θ_4, which is frequently encountered in the last position, is now in the first position in the ranking, and consequently, the other parameters are also in unusual positions. Model predictions and R^2 values are shown in Figure 3.15 and Table 3.22.

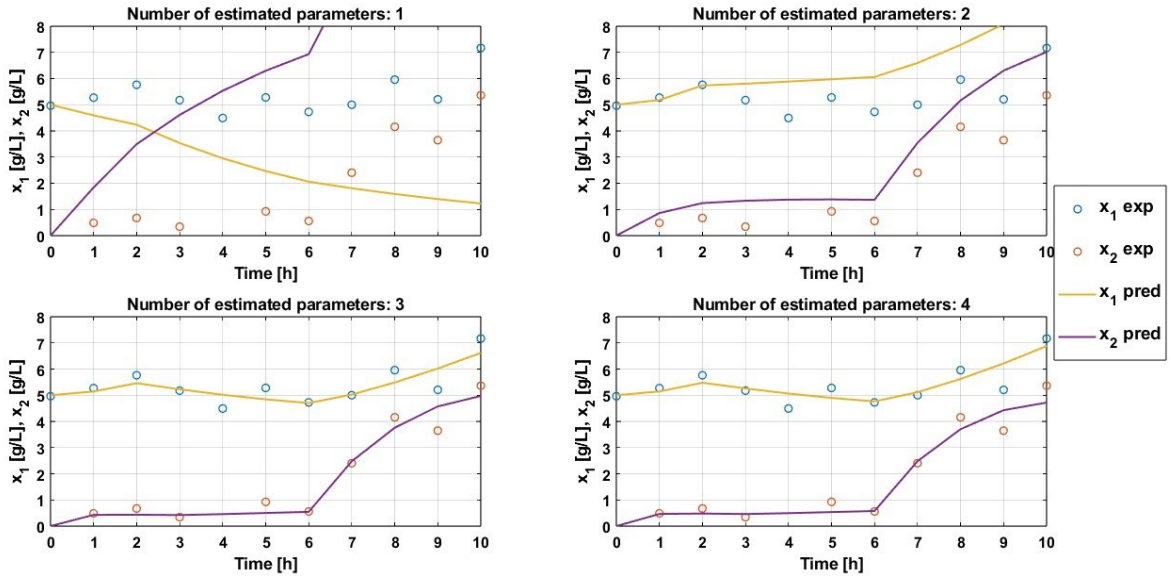


Figure 3.15. *Model predictions vs estimated parameters.*

Table 3.22. *R^2 values for each model output.*

Parameters estimated	x_1	x_2
1	-18.23	-13.90
2	-2.820	0.537
3	0.670	0.953

According to the results of the MSE-based method, only the first three ranked parameters should be estimated. Their estimates, along with the corresponding t-values are shown in Table 3.23.

Table 3.23. *Estimation of model parameters chosen by the MSE-based method.*

Parameter	*Initial guess*	*Estimate*	95% *CI*	95% *t-value*
θ_1	2.080×10^{-2}	0.278	0.180	1.540
θ_3	1.087	0.497	0.326	1.520
θ_4	3.310×10^{-2}	2.120×10^{-2}	0.133	0.159
				$t_{ref} = 1.729$

All the parameters t-value are lower than the reference one, indicating that the parameter estimates are not precise enough. At this point. One could either apply the MSE-based method again, with the parameter estimates obtained in the first iteration as initial guesses, and check if the new parameter estimates obtained are precise, or one could simply perform a MBDoE on the MSE-based method results of the first iteration, with the scope of improve the parameter estimates precision by designing a new informative experiment.

Table 3.24. shows the results of the MSE-based method when the parameter estimates obtained in the initial results (see Table 3.23) are used as initial guesses for a second iteration of the method. Note that parameter θ_2 was not selected for estimation previously, so its initial guess is kept at the one used in the first iteration.

Table 3.24. *Estimability results of the second iteration.*

Parameter	*Initial guess*	*Rank*	*Selected*
θ_1	0.278	1	yes
θ_2	0.232	3	no
θ_3	0.497	2	no
θ_4	2.120×10^{-2}	4	no

Using as initial guesses the parameter estimates obtained using the MSE-based method, in the second iteration of the method the parameter rankings changed, where the parameters retain their most common position, as seen in the Monte Carlo results. This time, however, the method selected only the first parameter for estimation, while keeping the others fixed. The model predictions with this set of initial guesses are shown in Figure 3.16 along with the R^2 values in Table 3.25.

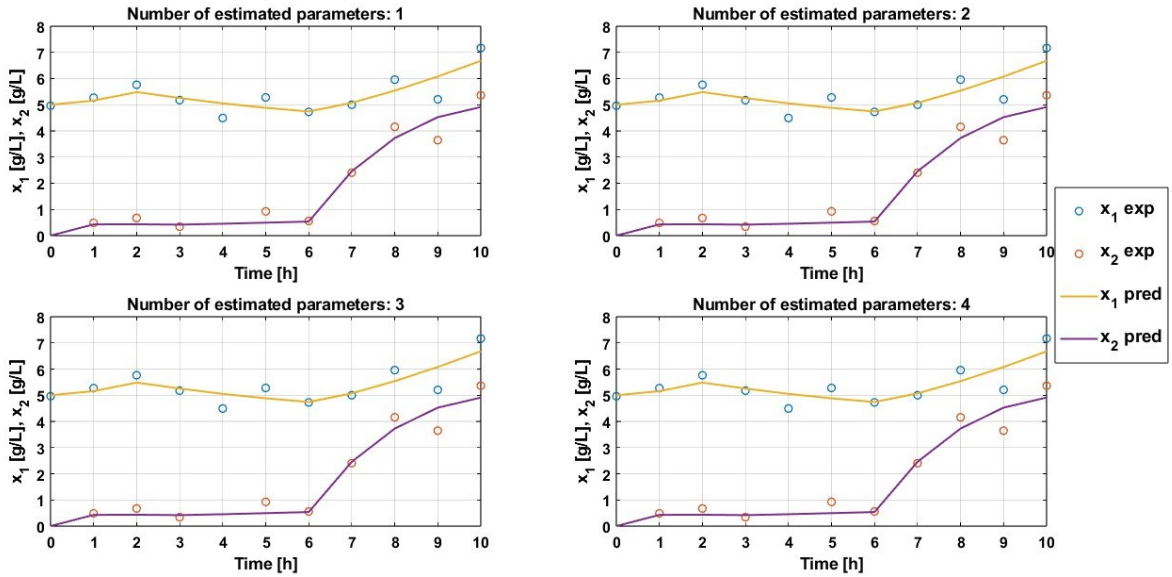


Figure 3.16. *Model predictions vs estimated parameters. Second iteration.*

Table 3.25. *R^2 values for each model output.*

Parameters estimated	x_1	x_2
1	0.670	0.953

While the estimated value of the first ranked parameter, which is the only one selected by the method for estimation, and its t-value is showed in Table 3.26. The R^2 value did not change when estimating more than one parameter.

Table 3.26. *Estimation of model parameters chosen by the MSE-based method. Second iteration.*

Parameter	***Initial guess***	***Estimate***	**95% *CI***	**95% *t-value***
θ_1	0.278	0.282	5.800×10^{-3}	48.83
				$t_{ref} = 1.720$

where parameter θ_1 has now a t-value larger than the reference one, indicating that the parameter estimate is precise enough in this second iteration.

Instead of repeating the MSE-based method with the parameter estimates found in Table 3.23 as initial guesses, one could use a MBDoE with the aim of improving the precision of the parameter estimates, by designing a new and more informative experiment. The values of the manipulate variables suggested by a MBDoE are shown in Table 3.27.

Table 3.27. *MBDoE results.*

	$u_1\ [h^{-1}]$	$u_2\ [g/L]$
$0 \leq t < 2\ h$	0.179	33.26
$2 \leq t < 6\ h$	0.140	25.01
$6 \leq t \leq 10\ h$	5.000×10^{-2}	27.87

Adding the new experiment in the parameter estimation task, and using the parameter estimates obtained in Table 3.23 as initial guesses, results in the following estimates:

Table 3.28. *Estimation of model parameters after MBDoE*

Parameter	***Initial guess***	***Estimate***	**95%** ***CI***	**95%** ***t-value***
θ_1	0.278	0.309	5.400×10^{-2}	5.700
θ_3	0.497	0.541	0.108	4.980
θ_4	2.120×10^{-2}	4.480×10^{-2}	2.300×10^{-2}	1.940
				$t_{ref} = 1.682$

where after a new informative experiment the parameters are estimated with enough precision, since all the parameters t-value are higher than the reference one.

3.7. Comparison with other estimability methods

Results of the estimability analysis obtained in §3.2 and §3.3 are compared to two methodologies used to assess the estimability of model parameters, which are the LSA and GSA, already mentioned in §2.1.3.

3.7.1. Local sensitivity analysis

The LSA is based on the computation of sensitivity coefficients S_{ij}, which are partial derivatives of the model predictions with respect to the model parameters. Sensitivity coefficients are calculated as follows:

$$S_{ij}(t_k) = \frac{\partial y_i(t_k)}{\partial \theta_j} \frac{s_{\theta j0}}{s_{yi}} \quad , \tag{3.3}$$

where the scaling factors $s_{\theta j0}$ and s_{yi} are needed so that the sensitivity coefficients are dimensionally consistent and comparable to each other. The LSA is performed when the uncertainty in the measures is high, and it is evaluated for both levels of uncertainty on parameter values used in the previous cases. Values of scaling factors and initial guesses are taken from §3.3. Figures 3.17 and 3.18 show the trend of the sensitivity coefficients with respect to time, for both levels of the uncertainty on parameter values.

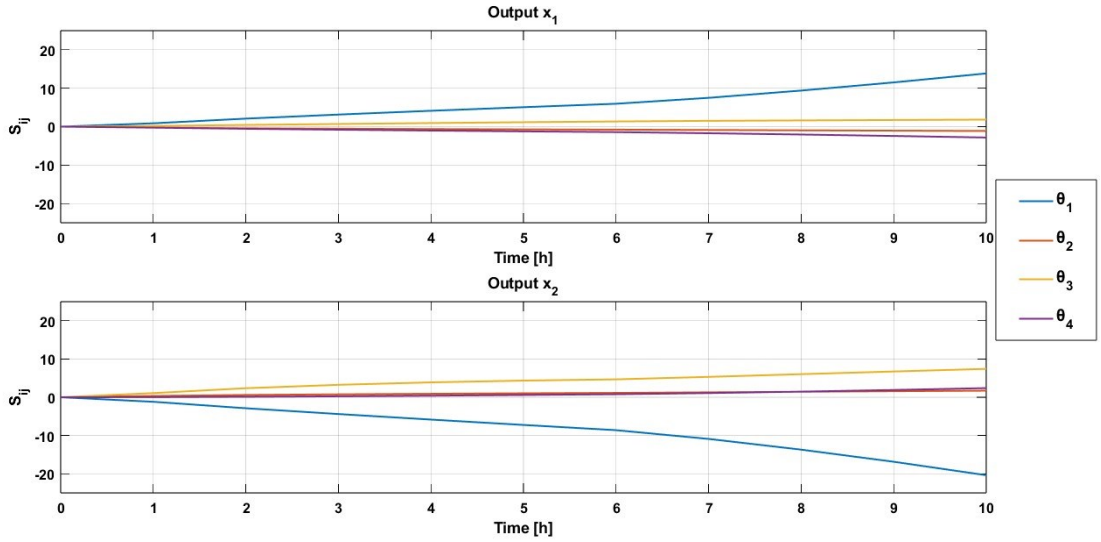


Figure 3.17. *LSA for the case of low uncertainty on parameter values.*

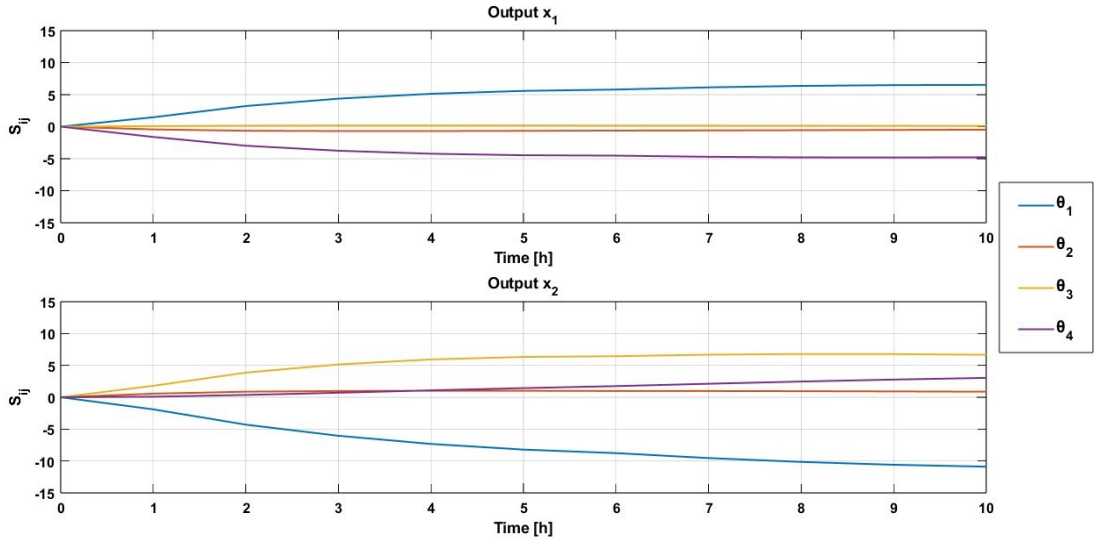


Figure 3.18. *LSA for the case of high uncertainty on parameter values.*

As seen in Figures 3.17 and 3.18, the model outputs x_1 and x_2 seem to be more sensible to perturbations of parameter θ_1, followed by parameter θ_3, which seems to be more important for output x_2. Parameter θ_4 seem to be influential for the output x_1 in the case of high parameter uncertainty, while parameter θ_2 seems to have little influence on the model outputs, having its values of sensitivity coefficients always close to zero.

From the LSA it can be extracted that parameter θ_1 is the most influential for the model outputs, while parameter θ_2 is the less influential, and can potentially be not estimable. The other remaining parameters exhibit some influence on model outputs, but no strong conclusions can be drawn. The LSA results on parameter θ_1 agree with the results of the orthogonalization method, since most of the time it selects θ_1 as the most estimable parameter, placing it in the top of the ranked list, while LSA results about θ_2 are somewhat compatible with the results of the MSE-based method since most of the time it does not select θ_2 to be estimated.

It is important to note that LSA depends on the initial guesses of model parameters, and that results are only valid close to the neighbourhood of the initial guesses, so no general conclusions can be drawn.

3.7.2. Global sensitivity analysis

The GSA is performed on the model of interest, and similarly to the LSA, it is performed for cases of low and high uncertainty on parameter values. The total effect indices $S_{T,i}$ are computed for every model parameter, every hour of the duration of the experiment. The results are shown in Figures 3.19 and 3.20.

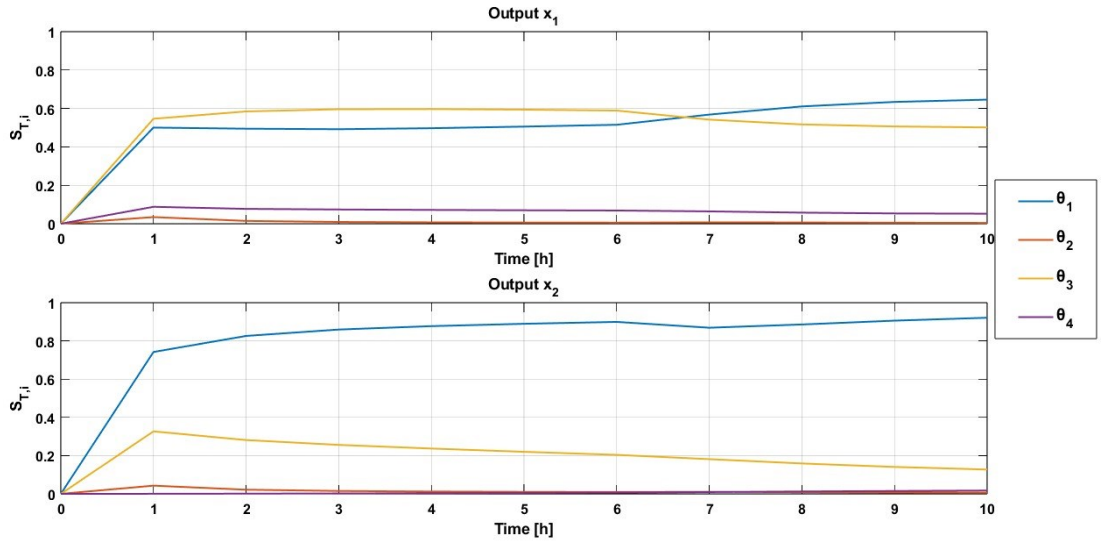


Figure 3.19. *Trend of the total effect index of model parameters. Case of low uncertainty on parameter values.*

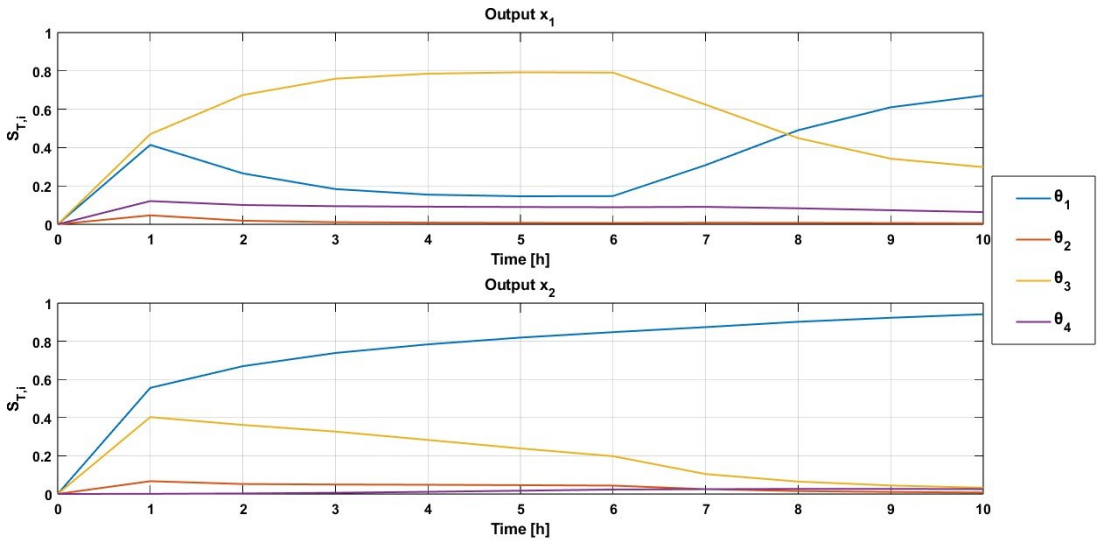


Figure 3.20. *Trend of the total effect index of model parameters. Case of high uncertainty on parameter values.*

Recall that $S_{T,i} = 0$ is a necessary and sufficient condition for an input factor (in this case, a parameter) to be non-influential (Saltelli, 2002). Therefore, parameters that exhibit this condition can be let free to vary over their range of uncertainty, without having significant contribution to the variance of the output.

GSA results show that for both cases of uncertainty, parameters θ_1 and θ_3 present high values of $S_{T,i}$, determining that these parameters are influential on the model outputs, and should be

preferably considered for the estimation task. Parameter θ_3, in particularly, seems to be more influential on the output x_1 and less influential on x_2. Parameters θ_2 and θ_4, instead, assume very low values of $S_{T,i}$ along the entire experiment, suggesting that their influence on the model output is not significant, and that they can potentially be fixed over their entire range of uncertainty, since estimability problems could arise, simplifying then the model and the parameter estimation task.

Overall, GSA results are in alignment with the orthogonalization and MSE-based method, since parameters θ_1 and θ_3 were selected most of the time by the orthogonalization method as the most influential parameters, and also selected by the MSE-based method to be estimated, while θ_2 and θ_4 were selected as the less influential ones and most of the time were excluded from the estimation task.

On the other hand, GSA does not consider the experimental data noise and does not give clear conditions to which parameters should be estimated.

3.7.3. Discussion

LSA-based ranking method is easy to implement, but cannot be recommended (i) when the relationship between model parameters and model predictions is highly nonlinear, (ii) when high uncertainty on parameter estimates is present, and (iii) when high parameters correlation is present. Moreover, LSA results only apply locally, i.e. near to the point in the parameter space at which the analysis is performed, so it is recommended when one has high certainty in the nominal parameter values. GSA on the other hand, has the advantage of taking into account the whole domain within the parameter space defined by the modeller, not depending on a specific choice of parameter values, therefore taking into account the parameter uncertainty. Moreover, GSA can also detect interactions between parameters. The main drawback of the GSA-based method – in particular, the variance-based Sobol's GSA – is the high computational burden, making its implementation almost impossible to computationally expensive models.

Either way, both methods give as a result a ranking of parameter influence on the model outputs, which in the case study presented, agree with each other, but lack of an objective result regarding on which parameters select for estimation, and which ones to fix at their nominal value. More specifically, GSA proposes a condition to fix parameter values when $S_{T,i} = 0$, but there is no clear condition when values of $S_{T,i}$ are for example slightly higher than zero. Moreover, both methods agree that parameters θ_1 and θ_3 are the most influential, which is also in line with the estimability method proposed by Wu et al., (2011).

Not only the estimability method proposed by Wu et al., (2011) has the capability of ranking model parameters as LSA and GSA, but also on pinpoints which parameter(s) one should select to be estimated, based on a mean-squared-error approach. Another advantage of the method is that it takes into account the available experimental data noise. Limitations of the method

proposed by Wu et al., (2011) are its dependence on initial parameter values (which is why a robustness test is recommended) and that the method considers the MSE for predictions at the experimental data for parameter estimation. Often, the modeler is not particularly interested in making predictions at the points where data is already available, but instead, accurate predictions may be desired at different sets of operating conditions where the model will be used (McLean and McAuley, 2012).

Chapter 4

Second case study: production of urethane

In this chapter, a more complex model is considered as a second instance to demonstrate the effectiveness of the estimability method proposed by Wu et al., (2011), considering different levels of measurement noise and uncertainty on parameter values. Then, the robustness of results is assessed.

4.1. Urethane production model

The kinetic model discussed by Galvanin et al., (2009) is considered as case study. The model describes the reaction of urethane in a simultaneous and consecutive set of reactions with chemical equilibrium:

$$\begin{aligned} A + B &\rightarrow C \\ A + C &\rightleftarrows D \\ 3A &\rightarrow E \end{aligned} \quad , \tag{4.1}$$

where A is phenylisocyanate, B is butanol, C is urethane, D is allophanate, and E is isocyanurate. Dimethylsulfoxide (S) is used as solvent. The experiments for these reactions are carried out in a semibatch reactor with two feed vessels v1 and v2, one for phenylisocyanate (and the solvent) and one for butanol (and the solvent). At the beginning, the reactor contains the solvent, phenylisocyanate and butanol. It is assumed that the reactor temperature T can be manipulated directly.

The model is represented by the following set of DAEs:

$$\begin{cases} \dfrac{dn_C}{dt} = V(r_1 - r_2 + r_3) \\ \dfrac{dn_D}{dt} = V(r_2 - r_3) \\ \dfrac{dn_E}{dt} = Vr_4 \\ n_A + n_C + 2n_D + 3n_E - n_A^0 - f^{v1}n_A^{v1} = 0 \\ n_B + n_C + n_D - n_B^0 - f^{v2}n_B^{v2} = 0 \\ n_S - n_S^0 - f^{v1}n_S^{v1} - f^{v2}n_S^{v2} = 0 \end{cases} , \tag{4.2}$$

with initial conditions for the three differential variables set to

$$n_C(0) = n_D(0) = n_E(0) = 0 \quad , \tag{4.3}$$

where the molar numbers n_i for species i are the state variables of the nonlinear DAE system. The following correlations also need to be considered:

$$\begin{cases} V = \sum_{i=1}^{6} \frac{n_i M_i}{\rho_i} \\ r_1 = k_1 \frac{n_A}{V}\frac{n_B}{V} = k_1 c_A c_B \\ r_2 = k_2 c_A c_C \\ r_3 = k_3 c_D \\ r_4 = k_4 c_A^2 \\ k_i = k_{ref,i}\, exp\left(-\frac{E_{a,i}}{R}\left(\frac{1}{T} - \frac{1}{T_{ref,i}}\right)\right), i = 1,2,4 \\ k_3 = \frac{k_2}{k_C} \\ k_C = k_{c2}\, exp\left(-\frac{\Delta h}{R}\left(\frac{1}{T} - \frac{1}{T_{g2}}\right)\right) \end{cases} . \tag{4.4}$$

In this model, eight parameters need to be estimated: the steric factors $k_{ref,i}$ ($i = 1,2,4$), the activation energies $E_{a,i}$ ($i = 1,2,4$), the equilibrium constant k_{c2}, and the reaction enthalpy Δh of the reversible reaction. The reaction rates r_i are expressed in [mol/(L · h)] and the volume V in [L]. Molar masses M_i [kg/mol], densities ρ_i [kg/m^3], reference temperatures $T_{ref,1}$, $T_{ref,2}$, $T_{ref,4}$, T_{g2} [K], and the gas constant R [J/(mol · K)] are set as constant, as shown in Table 4.1. Values of the constants are taken from Bauer et al., (2000).

Table 4.7. *Constants in the model for the reaction of urethane.*

Molar masses **[kg/mol]**	***Densities*** **[kg/m^3]**	***Reference temperatures*** **[K]**
$M_A = 0.119$	$\rho_A = 1.095 \times 10^3$	$T_{ref,1} = 363.16$
$M_B = 7.412 \times 10^{-2}$	$\rho_B = 809.00$	$T_{ref,2} = 363.16$
$M_C = 0.193$	$\rho_C = 1.415 \times 10^3$	$T_{ref,4} = 363.16$
$M_D = 0.312$	$\rho_D = 1.528 \times 10^3$	$T_{g2} = 363.16$
$M_E = 0.357$	$\rho_E = 1.451 \times 10^3$	***Gas constant*** **[J/(mol·K)]**
$M_S = 7.806 \times 10^{-2}$	$\rho_S = 1.101 \times 10^3$	$R = 8.314$

Feeds f^{v1} and f^{v2} are the "accumulated" feeds from the two feed vessels. They can vary from zero up to the initial molar hold-ups of the vessels, described by n_A^{v1}, n_B^{v2}, n_S^{v1}, and n_S^{v2}, i.e., the initial molar numbers of phenylisocyanate, butanol, and solvent within the two feed vessels.

Feeds f^{v1} and f^{v2} and temperature T are the experiment design variables. It is assumed that the temperature profile T is discretized as a piecewise linear function within the range of 300 – 473 K, and that the total duration of the experiment is 90 h. Figure 4.1 shows how the temperature profile was manipulated in this study.

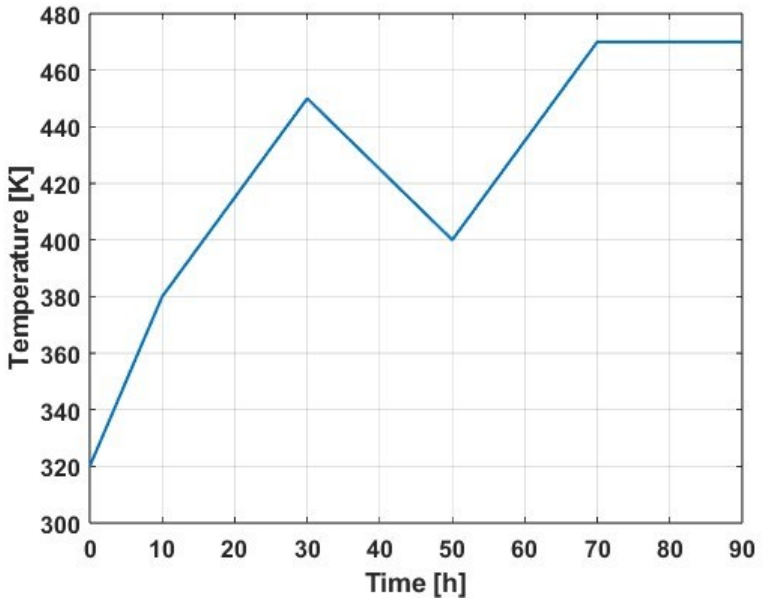


Figure 4.1. *Temperature profile.*

The measured variables are the molar concentrations [mol/L] of urethane (c_C), of allophanate (c_D), and of isocyanurate (c_E). During a single experiment, samples are taken every hour, for the total duration of the experiment. The experiment conditions are reported in Table 4.2.

Table 4.8. *Process settings at nominal conditions.*

Variable	***Units***	***Nominal value***
t	h	90.00
$n_A(0)$	mol	1.000×10^3
$n_B(0)$	mol	50.00
$n_S(0)$	mol	10.00
n_A^{v1}	mol	0.00
n_B^{v2}	mol	0.00
n_S^{v1}	mol	0.00
n_S^{v2}	mol	0.00

The MSE-based method proposed by Wu et al., (2011) combined with the orthogonalization method is then applied to the process of urethane, and like the case of the fermentation process

seen in Chapter 3, different levels of measurement noise and uncertainty on parameter values are considered.

4.2. Methodology results with low experimental data noise

A single measurement sensor is present in the process, which in the case of low experimental data noise, it is assumed that the sensor measures the measured variables c_C, c_D, and c_E with a constant standard error of $\sigma = 4.000\times10^{-3}$ mol/L. Therefore, the scaling factor that represent the uncertainty of the measures is set as $s_{yi} = \sigma = 4.000\times10^{-3}$ mol/L. As in chapter 3, the methodology is assessed for cases of low and high uncertainty on parameter values.

4.2.1. Low uncertainty on parameter values

The scaling factor that reflects the uncertainty on parameter values is calculated as Equation (3.2), but differently from the fermentation process, low uncertainty corresponds to a range of $\pm$20% of the "true" parameter values, thus $lb_{\theta j} = \theta_j - (0.200 \times \theta_j)$ and $ub_{\theta j} = \theta_j + (0.200 \times \theta_j)$ are set as lower and upper values of the uncertainty range. Parameter initial guesses, along with the parameter "true" values and uncertainty factors are presented in Table 4.3.

Table 4.9. *"True" parameter values, initial parameter guesses and uncertainty factors.*

Parameter	"True" value	Initial guess	Uncertainty $s_{\theta j0}$
$k_{ref,1}$ [L/mol · h]	1.250×10^{-3}	1.100×10^{-3}	4.400×10^{-4}
$k_{ref,2}$ [L/mol · h]	7.290×10^{-6}	8.700×10^{-6}	3.480×10^{-6}
$k_{ref,4}$ [L/mol · h]	8.800×10^{-7}	1.000×10^{-6}	4.000×10^{-7}
$E_{a,1}$ [J/mol]	2.944×10^{4}	2.500×10^{4}	1.000×10^{4}
$E_{a,2}$ [J/mol]	7.101×10^{4}	6.000×10^{4}	2.400×10^{4}
$E_{a,4}$ [J/mol]	2.302×10^{4}	1.900×10^{4}	7.600×10^{3}
k_{c2} [L/mol]	0.217	0.180	7.200×10^{-2}
Δh [J/mol]	-1.830×10^{4}	-2.000×10^{4}	8.000×10^{3}

The results of the methodology are shown in Table 4.4, along with the trends of the objective function J and corrected critical ratio $r_{CC,k}$, shown in Figure 4.2.

Table 4.10. *Estimability analysis results. Case of low experimental noise and low uncertainty on parameter values.*

Parameter	Rank	Selected
$k_{ref,1}$	3	yes
$k_{ref,2}$	5	yes
$k_{ref,4}$	4	yes
$E_{a,1}$	1	yes
$E_{a,2}$	2	yes
$E_{a,4}$	8	no
k_{c2}	7	no
Δh	6	yes

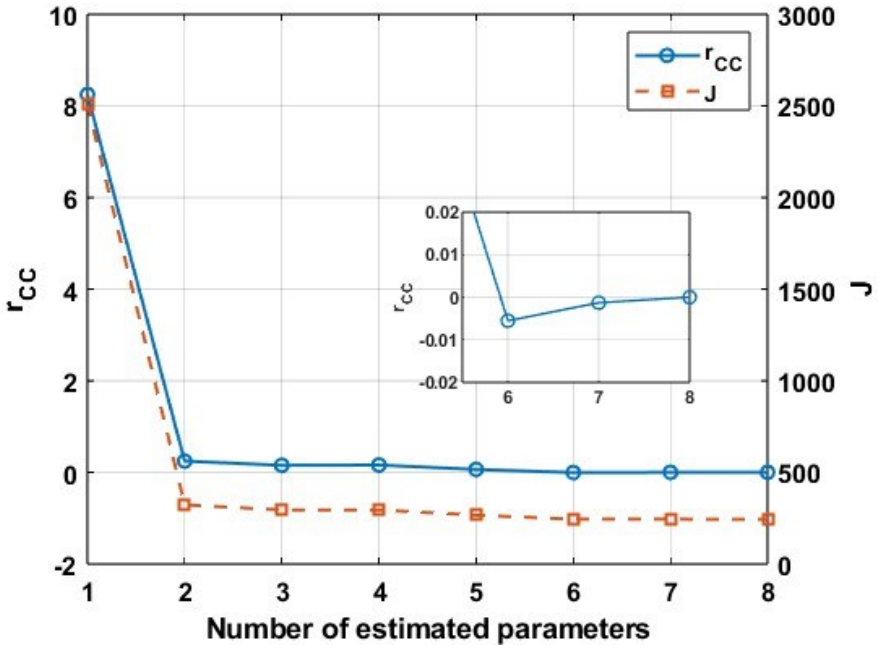


Figure 4.2. *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Case of low noise and low uncertainty on parameter values.*

where the first six top ranked parameters are selected to be estimated. Moreover, the objective function value does not improve significantly when more than two parameters are estimated. Model predictions for every subset of estimated parameters compared to the experimental data are presented in Figure 4.3, while the R^2 values, which assess the goodness of the data fit, are shown in Table 4.5. Note that all estimated model parameters are statistically satisfactory (t-values higher than the reference), as reported in Table 4.6.

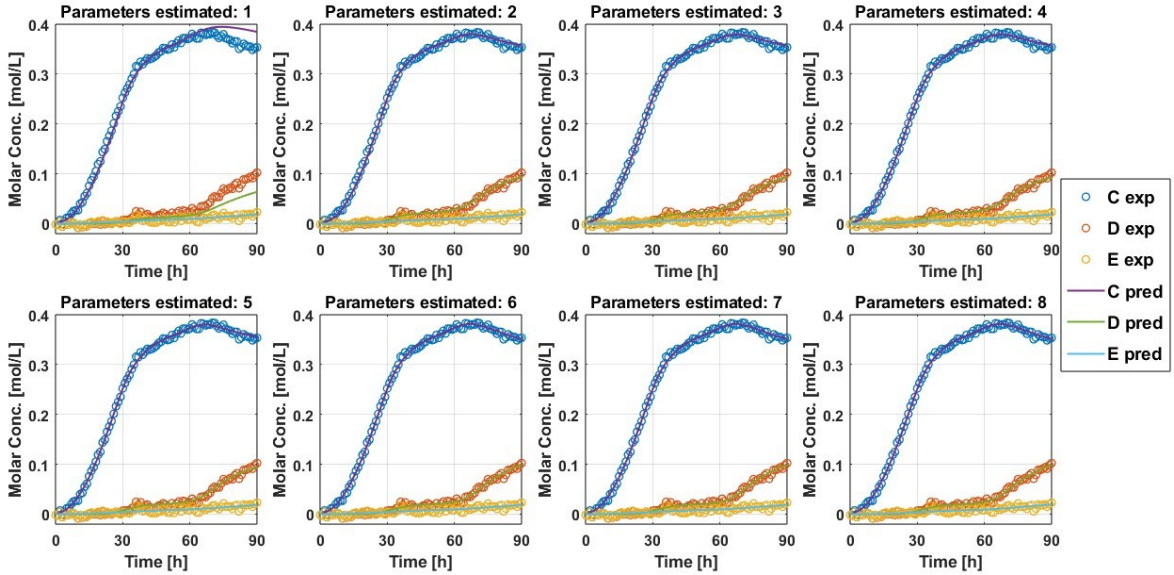


Figure 4.3. *Model predictions of every subset of estimated parameters vs experimental data. Case of low noise and low uncertainty on parameter values.*

Table 4.5. *R^2 values for each model output. Case of low experimental noise and low uncertainty on parameter values.*

Parameters estimated	c_C	c_D	c_E
1	0.988	0.761	0.710
2	0.998	0.977	0.710
3	0.999	0.977	0.710
4	0.999	0.978	0.710
5	0.999	0.981	0.710
6	0.999	0.984	0.710

Table 4.6. *Precision of parameter estimates of the estimability analysis. Case of low experimental noise and low uncertainty on parameter values.*

Parameter	***"True" value***	***Initial guess***	***Estimate***	***95% t-value***
$k_{ref,1}$	1.250×10^{-3}	1.100×10^{-3}	1.200×10^{-3}	7.619
$k_{ref,2}$	7.290×10^{-6}	8.700×10^{-6}	6.956×10^{-6}	62.94
$k_{ref,4}$	8.800×10^{-7}	1.000×10^{-6}	9.994×10^{-7}	5.533
$E_{a,1}$	2.944×10^{4}	2.500×10^{4}	2.970×10^{4}	11.08
$E_{a,2}$	7.101×10^{4}	6.000×10^{4}	7.182×10^{4}	105.81
Δh	-1.830×10^{4}	-2.000×10^{4}	-1.600×10^{4}	17.00
				$t_{ref} = 1.650$

Model predictions seem satisfactory even for a small number of parameters selected for estimation, in fact, R^2 values do not change significantly when two or more parameters are estimated. However, the MSE-based method still selected the top six parameters to be estimated.

4.2.2. High uncertainty on parameter values

It is assumed that when the uncertainty on parameter values is high, the parameters can range from $\pm$50% of the parameters "true" value. Therefore, scaling factors that reflect the uncertainty are calculated with lower and upper values $lb_{\theta j} = \theta_j - (0.500 \times \theta_j)$ and $ub_{\theta j} = \theta_j + (0.500 \times \theta_j)$. Parameter initial guesses, along with the parameter "true" values and uncertainty factors are presented in Table 4.7.

Table 4.7. *"True" parameter values, initial parameter guesses and uncertainty factors.*

Parameter	"True" value	Initial guess	Uncertainty $s_{\theta j0}$
$k_{ref,1}$ [L/mol · h]	1.250×10^{-3}	1.850×10^{-3}	1.850×10^{-3}
$k_{ref,2}$ [L/mol · h]	7.290×10^{-6}	3.700×10^{-6}	3.700×10^{-6}
$k_{ref,4}$ [L/mol · h]	8.800×10^{-7}	1.300×10^{-6}	1.300×10^{-6}
$E_{a,1}$ [J/mol]	2.944×10^{4}	1.500×10^{4}	1.500×10^{4}
$E_{a,2}$ [J/mol]	7.101×10^{4}	3.700×10^{4}	3.700×10^{4}
$E_{a,4}$ [J/mol]	2.302×10^{4}	3.300×10^{4}	3.300×10^{4}
k_{c2} [L/mol]	0.217	0.110	0.110
Δh [J/mol]	-1.830×10^{4}	-1.000×10^{4}	1.000×10^{4}

The results of the methodology are presented in Table 4.8, while trend of J and $r_{CC,k}$ are shown in Figure 4.4.

Table 4.8. *Estimability analysis results. Case of low experimental noise and high uncertainty on parameter values.*

Parameter	Rank	Selected
$k_{ref,1}$	1	yes
$k_{ref,2}$	6	yes
$k_{ref,4}$	5	yes
$E_{a,1}$	3	yes
$E_{a,2}$	4	yes
$E_{a,4}$	2	yes
k_{c2}	7	no
Δh	8	no

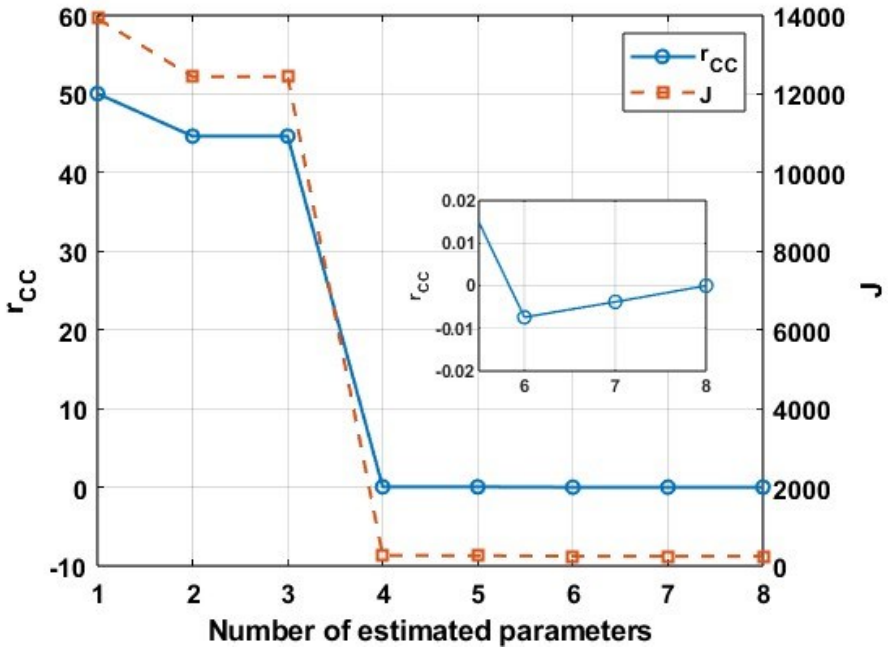


Figure 4.4. *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Case of low noise and high uncertainty on parameter values.*

Some parameters changed their rankings significantly ($k_{ref,1}$, $E_{a,4}$), while others maintained or slightly changed their positions. The MSE-based method selected the top six ranked parameters to be estimated, and the objective function value did not change significantly when four or more parameters are estimated. Model predictions for every subset of estimated parameters compared to the experimental data are presented in Figure 4.5, and the corresponding R^2 in Table 4.9.

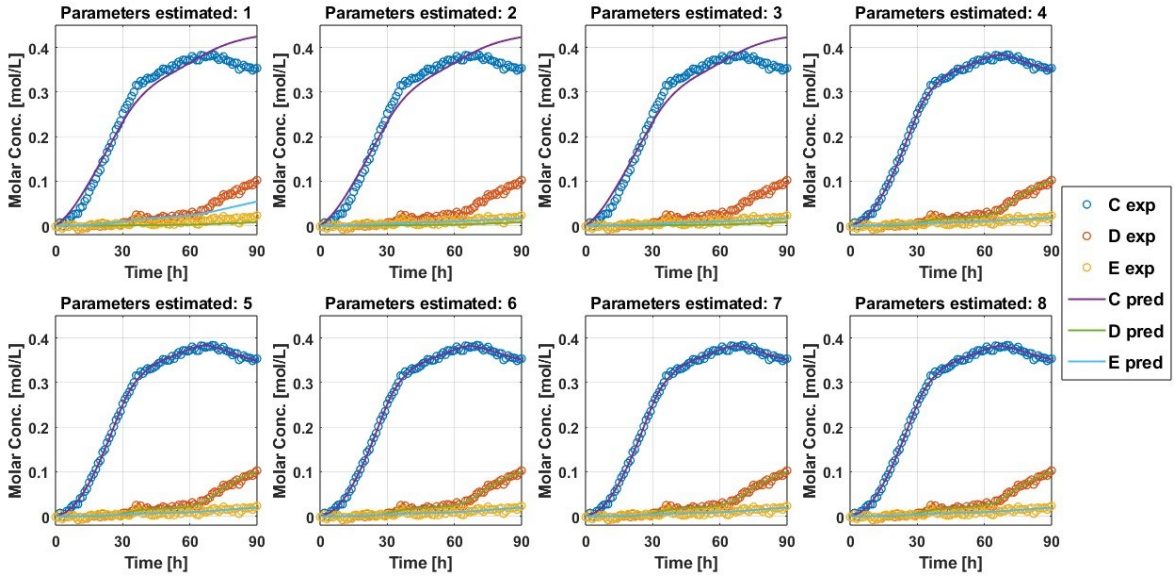


Figure 4.5. *Model predictions of every subset of estimated parameters vs experimental data. Case of low noise and high uncertainty on parameter values.*

Table 4.9. *R^2 values for each model output. Case of low experimental noise and high uncertainty on parameter values.*

Parameters estimated	c_C	c_D	c_E
1	0.952	-0.506	-4.218
2	0.953	-0.506	0.699
3	0.953	-0.506	0.698
4	0.999	0.982	0.703
5	0.999	0.982	0.714
6	0.999	0.984	0.714

Table 4.10. *Precision of parameter estimates of the estimability analysis. Case of low experimental noise and high uncertainty on parameter values.*

Parameter	***"True" value***	***Initial guess***	***Estimate***	***95% t-value***
$k_{ref,1}$	1.250×10^{-3}	1.850×10^{-3}	1.300×10^{-3}	31.61
$k_{ref,2}$	7.290×10^{-6}	3.700×10^{-6}	6.080×10^{-6}	28.49
$k_{ref,4}$	8.800×10^{-7}	1.300×10^{-6}	1.176×10^{-6}	9.454
$E_{a,1}$	2.944×10^{4}	1.500×10^{4}	2.933×10^{4}	39.01
$E_{a,2}$	7.101×10^{4}	3.700×10^{4}	7.408×10^{4}	81.75
$E_{a,4}$	2.302×10^{4}	3.300×10^{4}	1.637×10^{4}	3.569
				$t_{ref} = 1.650$

where confirming the trend of J, model predictions seem to not change significantly after having estimated the top four ranked parameters. Like the previous case, all estimated model parameters are statistically satisfactory, as shown in Table 4.10.

4.3. Methodology results with high experimental data noise

When the experimental data noise is high, it is assumed that the sensor measures the measured variables c_C, c_D, and c_E with a constant standard error of $\sigma = 1.000 \times 10^{-2}$ mol/L. Therefore, the scaling factor that considers measurement uncertainty is set as $s_{yi} = \sigma = 1.000 \times 10^{-2}$ mol/L. Cases of low and high uncertainty on parameter values are considered next.

4.3.1. Low uncertainty on parameter values

The initial parameter guesses used are the same ones from the case of low experimental noise and uncertainty (refer to Table 4.3.). The methodology results are shown in Table 4.11 and trends of J and $r_{CC,k}$ in Figure 4.6.

Table 4.11. *Estimability analysis results. Case of high experimental noise and low uncertainty on parameter values.*

Parameter	Rank	Selected
$k_{ref,1}$	4	no
$k_{ref,2}$	6	no
$k_{ref,4}$	5	no
$E_{a,1}$	2	yes
$E_{a,2}$	1	yes
$E_{a,4}$	7	no
k_{c2}	8	no
Δh	3	yes

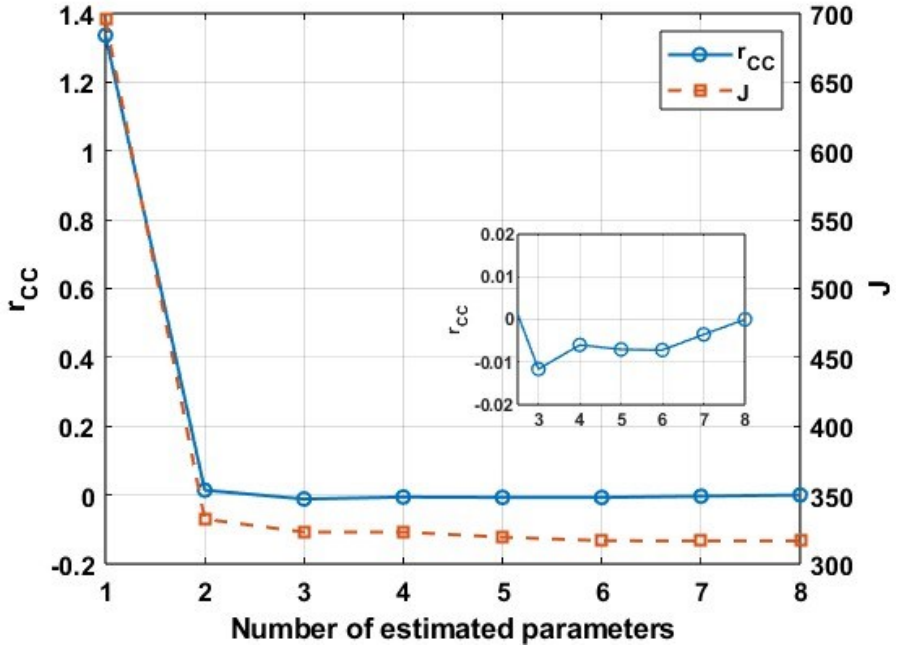


Figure 4.6. *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Case of high noise and low uncertainty on parameter values.*

where the estimability method selects the top three ranked parameters to be estimated. Model predictions for every subset of estimated parameters compared to the experimental data are presented in Figure 4.7, along with the R^2 values in Table 4.12. Note that parameter Δh estimate resulted to not be statistically satisfactory (t-value lower than the reference one), as reported in Table 4.13. However, even though only two model parameters are estimated with a statistically satisfactory precision (while one is not precise statistically and the rest are not estimated, thus

fixed to their initial guess), the model still gives satisfactory predictions (see R^2 values in Table 4.12).

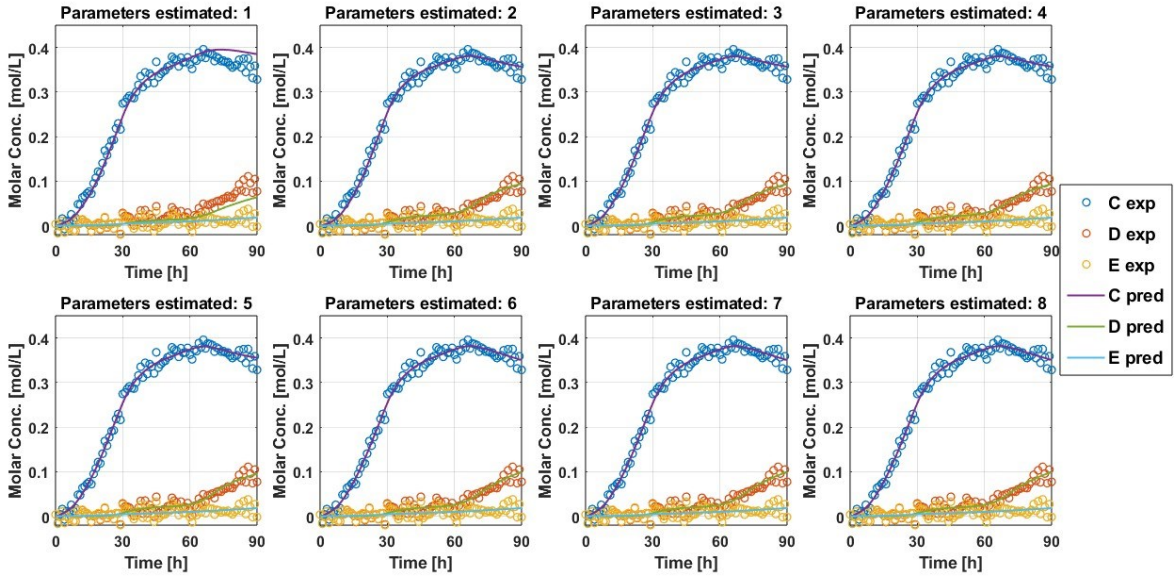


Figure 4.7. *Model predictions of every subset of estimated parameters vs experimental data. Case of high noise and low uncertainty on parameter values.*

Table 4.12. *R^2 values for each model output. Case of high experimental noise and low uncertainty on parameter values.*

Parameters estimated	c_C	c_D	c_E
1	0.983	0.688	0.245
2	0.994	0.906	0.245
3	0.994	0.906	0.245

Table 4.13. *Precision of parameter estimates of the estimability analysis. Case of high experimental noise and low uncertainty on parameter values. * = estimated model parameter is not statistically satisfactory (i.e., t-value lower than the reference).*

Parameter	***"True" value***	***Initial guess***	***Estimate***	***95% t-value***
$E_{a,1}$	2.944×10^4	2.500×10^4	2.779×10^4	11.46
$E_{a,2}$	7.101×10^4	6.000×10^4	6.880×10^4	4.244
Δh	-1.830×10^4	-2.000×10^4	-1.464×10^4	1.241*
				$t_{ref} = 1.650$

Model predictions do not change significantly when two of more parameters are estimated, despite suggesting that predictions for the molar concentration of specie E are not very accurate

(low R^2 value). This is likely due to the high measurement noise of the system under investigation, which has the same magnitude of the values of E concentration. In other words, the high noise makes R^2 values untrustworthy when referring to specie E. Therefore, overall model performance is good (see R^2 values of species C and D).

4.3.2. High uncertainty on parameter values

Parameter initial guesses and uncertainty used in this case are the same as the analog case where the experimental noise is low (refer to Table 4.7). The estimability results are shown in Table 4.14 and trends of J and $r_{CC,k}$ in Figure 4.8.

Table 4.14. *Estimability analysis results. Case of high experimental noise and high uncertainty on parameter values.*

Parameter	Rank	Selected
$k_{ref,1}$	1	yes
$k_{ref,2}$	6	no
$k_{ref,4}$	5	no
$E_{a,1}$	3	yes
$E_{a,2}$	4	yes
$E_{a,4}$	2	yes
k_{c2}	7	no
Δh	8	no

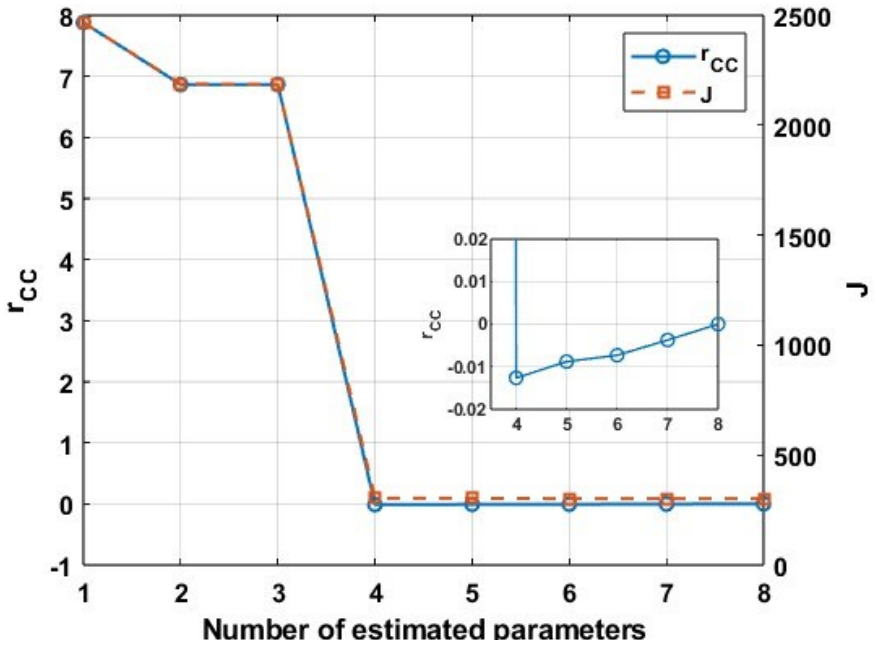


Figure 4.8. *Effect of the number of parameters estimated on the objective function J and $r_{CC,k}$ values obtained using the algorithm in Table 2.2. Case of high noise and high uncertainty on parameter values.*

where this time, the estimability method selects the first four parameters to be estimated. Parameters ranking, however, remained the same as the analog case with low experimental noise. Model predictions for every subset of estimated parameters compared to the experimental data are presented in Figure 4.9, and R^2 values are shown in Table 4.15. Similar comments to observations drawn in the previous section regarding R^2 values with reference to specie E can be done here.

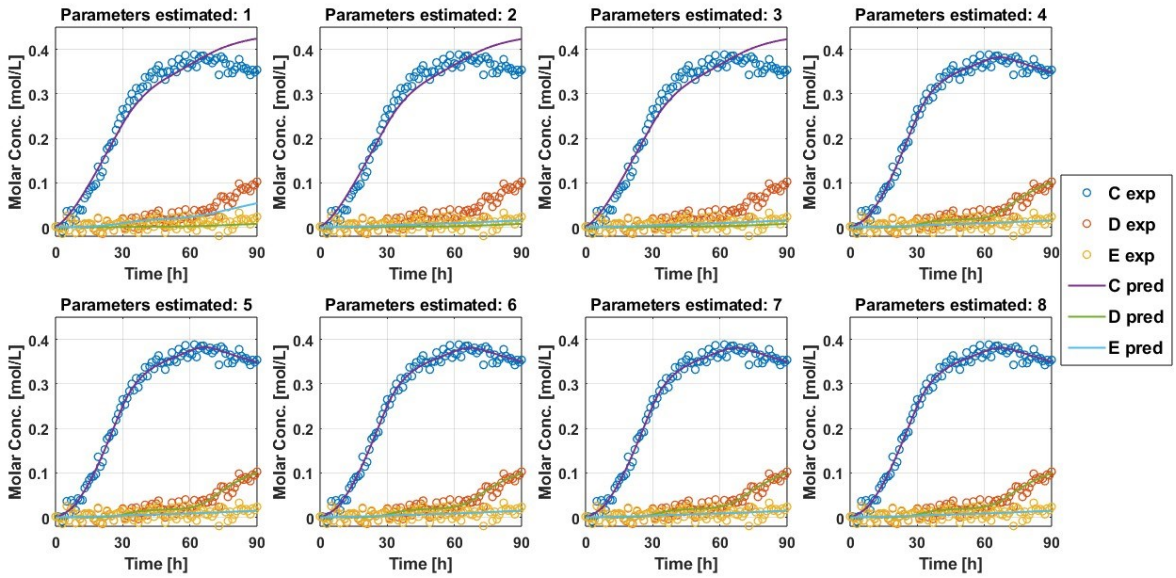


Figure 4.9. *Model predictions of every subset of estimated parameters vs experimental data. Case of high noise and high uncertainty on parameter values.*

Table 4.15. *R^2 values for each model output. Case of high experimental noise and high uncertainty on parameter values.*

Parameters estimated	c_C	c_D	c_E
1	0.944	-0.515	-2.195
2	0.945	-0.514	0.121
3	0.945	-0.514	0.121
4	0.993	0.873	0.121

Table 4.16. *Precision of parameter estimates of the estimability analysis. Case of high experimental noise and high uncertainty on parameter values.*

Parameter	***"True" value***	***Initial guess***	***Estimate***	***95% t-value***
$k_{ref,1}$	1.250×10^{-3}	1.850×10^{-3}	1.300×10^{-3}	31.61
$E_{a,1}$	2.944×10^{4}	1.500×10^{4}	2.752×10^{4}	14.63
$E_{a,2}$	7.101×10^{4}	3.700×10^{4}	8.143×10^{4}	16.70
$E_{a,4}$	2.302×10^{4}	3.300×10^{4}	1.151×10^{4}	4.141
				$t_{ref} = 1.650$

As seen by the trend of J, model predictions do not change significantly when four or more parameters are estimated. Moreover, all estimated model parameters are statistically satisfactory, as shown in Table 4.16.

4.4. Discussion

The estimability analysis results are summarized in Table 4.17.

Table 4.17. *Comparison between all the estimability results.*

Parameter	***Low noise***				***High noise***			
	Low uncertainty on parameter values		High uncertainty on parameter values		Low uncertainty on parameter values		High uncertainty on parameter values	
	Rank	Selected	Rank	Selected	Rank	Selected	Rank	Selected
$k_{ref,1}$	3	yes	1	yes	4	no	1	yes
$k_{ref,2}$	5	yes	6	yes	6	no	6	no
$k_{ref,4}$	4	yes	5	yes	5	no	5	no
$E_{a,1}$	1	yes	3	yes	2	yes	3	yes
$E_{a,2}$	2	yes	4	yes	1	yes	4	yes
$E_{a,4}$	8	no	2	yes	7	no	2	yes
k_{c2}	7	no	7	no	8	no	7	no
Δh	6	yes	8	no	3	yes	8	no

The estimability method tends to select more parameters to be estimated when the experimental data noise is high (i.e., three parameters need to be estimated when the uncertainty on parameter values is low, four parameters need to be estimated when uncertainty on parameter values is high). Conversely, when the experimental data noise is low, the method selects for both cases of uncertainty six parameters to be estimated. This trend of selecting less parameters to be

estimated when the experimental data noise is high is also found in the previous Chapter. Not only the number of parameters selected is affected by the measurement noise, but also on the initial values of parameter guesses, since when the values of parameter guesses are not close to the "true" values (case of high uncertainty on parameter values), the method tends to select more parameters to be estimated (see when the noise is high).

Parameters ranking remains the same when the uncertainty on parameter values is high, independently of the noise, while it slightly changes when the uncertainty on parameter values is low. When all cases are analyzed, some parameters are commonly found in the top positions (i.e. $E_{a,1}$ and $E_{a,2}$), while some in the last positions (i.e. k_{c2} and Δh), but no absolute conclusions can be drawn about the ranking of influence of the parameters.

4.5. Robustness of the results

As in Chapter 3, the robustness of the results is assessed using a Monte Carlo approach, where 100 different sets of parameter initial guesses are taken from the range of uncertainty on parameter values assumed.

4.5.1. Results with low experimental data noise

The frequency of the rank for each parameter is shown in the shaded box diagram in Figure 4.10.

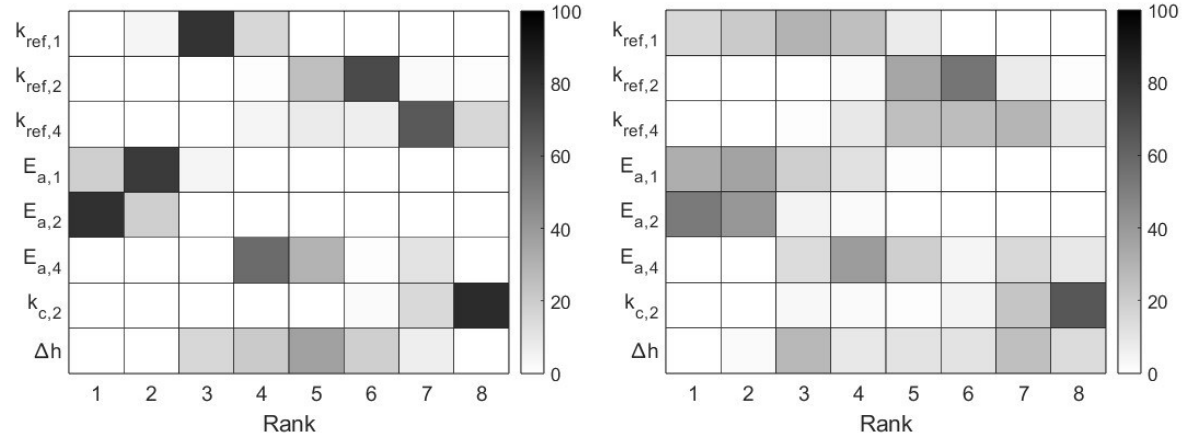


Figure 4.140. *Frequency of parameters rankings obtained using the algorithm in Table 2.1 with 100 random initial parameter guesses. Case of low uncertainty (a), and high uncertainty (b) on parameter values.*

The degree of shading for each cell depends on the frequency of that rank for the parameter. For example, in the case of low uncertainty on parameter values, parameter $k_{ref,1}$, which is ranked third in Table 4.4, ranks as third using approximately 80 out of 100 sets of initial parameter guesses, while parameter $E_{a,2}$, which is ranked second in Table 4.4, results to rank as the top parameter more frequently (nearly 80% of the time). This inversion also happens with parameter $E_{a,1}$, whereas the ranks for other parameters are more variable.

Rankings when the uncertainty on parameter values is low are more defined (a lot of positions are well established), while rankings when the uncertainty on parameter values is high are more variable and so less certain. Despite that, there is a tendency of parameters $E_{a,1}$, $E_{a,2}$ and $k_{ref,1}$ to be at the top of the ranked list, while k_{c2}, Δh and $k_{ref,4}$ are more commonly encountered in the bottom of the ranked list.

The frequencies of the optimal number of parameters selected for estimation are shown in Figure 4.11.

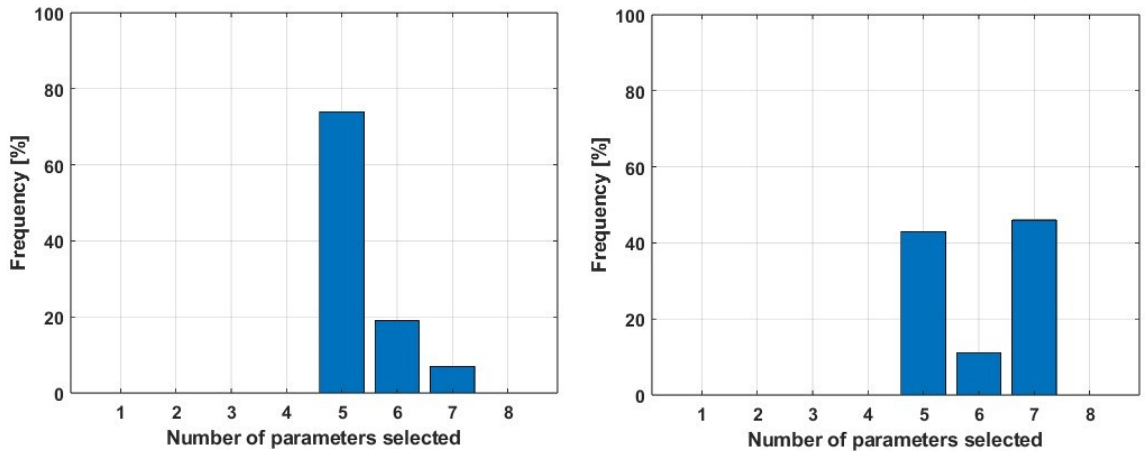
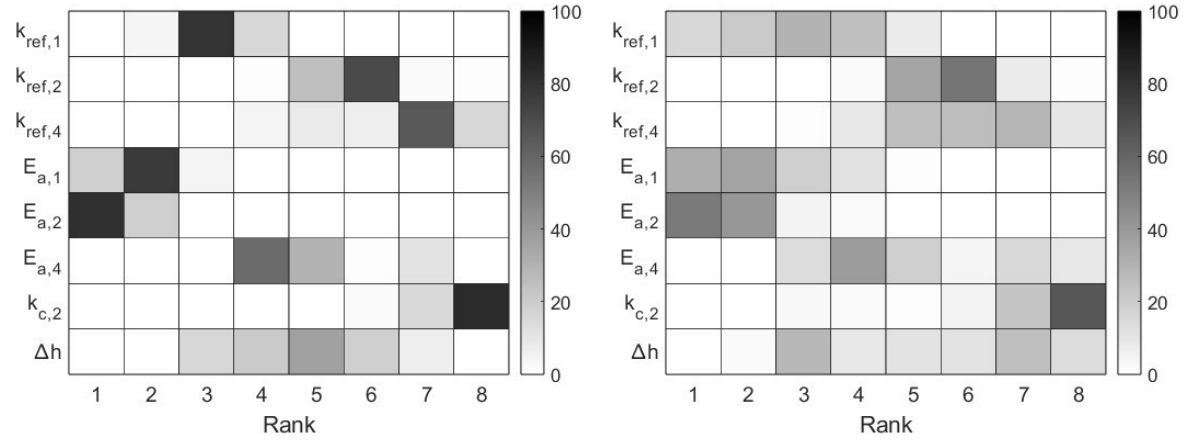


Figure 4.11. *Frequency of the number of parameters selected obtained using the algorithm in Table 2.2 with 100 random initial parameter guesses. Case of low uncertainty (a) and high uncertainty (b) on parameter values.*

When the uncertainty on parameter values is low, most of the time the top five parameters are selected for estimation (nearly 80% of the time), while when the uncertainty on parameter values is high, the method tends to select five or seven parameters to be estimated, with both choices selected approximately 40% of the time.

4.5.2. Results with high experimental data noise

The frequency of the rank for each parameter is shown in the shaded box diagram in Figure 4.12.

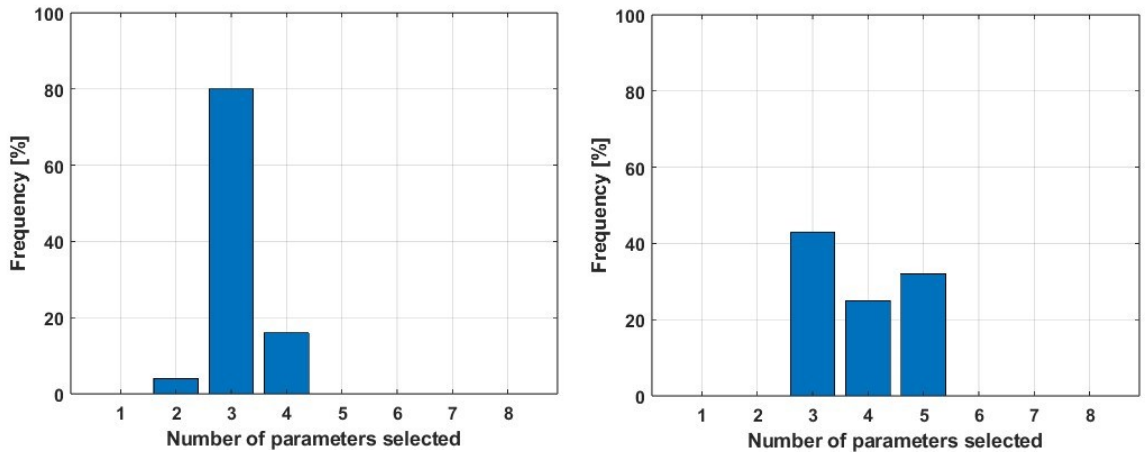


(a) **(b)**

Figure 4.12. *Frequency of parameters rankings obtained using the algorithm in Table 2.1 with 100 random initial parameter guesses. Case of low uncertainty (a), and high uncertainty (b) on parameter values.*

As the case of low measurement noise, when the uncertainty on parameter values is low, parameter rankings are more regular, where parameters tend to retain their ranking most of the time, while the ranking is more variable when the uncertainty on parameter values is high. In both cases, the rankings obtained in §4.3 agree for the most part with the most common positions found in the robustness test.

The frequencies of the optimal number of parameters selected for estimation are shown in Figure 4.13.



(a) **(b)**

Figure 4.13. *Frequency of the number of parameters selected obtained using the algorithm in Table 2.2 with 100 random initial parameter guesses. Case of low uncertainty (a) and high uncertainty (b) on parameter values.*

When the uncertainty on parameter values is low, the method selects for nearly 80% of the time three parameters to be estimated, which is also the number obtained in §4.3.1, while when the uncertainty on parameter values is high, the method also selects with a higher frequency three parameters to be estimated, but four and five parameters are also selected with a similar frequency.

4.5.3. Discussion

The robustness test showed that the estimability method tends to select a lower number of parameters to be estimated from the ranked list (i.e. it selected three parameters to be estimated with a higher frequency, for both cases of uncertainty on parameter values) when the experimental data noise is low, while the method tends to select a higher number of parameters to be estimated when the experimental data noise is high (5 and 7 parameters are selected with a higher frequency when the uncertainty on parameter values is low and high, respectively).
Moreover, when the uncertainty on parameter values is lower, the method selected the most common outcome of the MSE-based method with a higher frequency (an optimal number of parameters selected is picked approximately 80 out of 100 times), while when the uncertainty on parameter values is higher, there was not a significant conclusive result on the number of parameters selected (for example, five and seven parameters are selected with almost the same frequency when the experimental noise is low).
Meanwhile, parameter rankings appear not to change through the different levels of experimental noise; conversely, they change for different levels of uncertainty on parameter values. However, while it is true that when the uncertainty on parameter values is higher, the rankings are more variable, it appears that for both cases of uncertainty, some parameters are more commonly picked to be at the top of the ranked list ($E_{a,1}$, $E_{a,2}$, $k_{ref,1}$), some in the middle ($E_{a,4}$, $k_{ref,2}$), and some in the bottom (k_{c2}, $k_{ref,4}$, Δh).

Conclusions

The estimability method proposed by Wu et al., (2011) was applied to two case studies: (i) a fermentation process where the mathematical model describing the process has 4 model parameters, and (ii) a urethane production process with 8 model parameters. The method was then tested considering that (i) the available experimental data are affected by different levels of measurement noise, and that (ii) initial parameter guesses are characterized by different levels of uncertainty. Based on the presented case studies, some general observations can be drawn:

- When the measurement noise is higher, the method tends to select less parameters to be estimated compared to when the measurement noise is lower;
- When the uncertainty on parameter values is higher (i.e. when the parameter initial guesses are distant from the "true" values), the method tends to select more parameters to be estimated compared to when the uncertainty on parameter values is lower.

A robustness test based on a Monte Carlo approach was then applied to overcome the issue of high dependency on initial parameter guesses, and its results mostly confirmed what was previously found. However, if the initial parameter guesses are very different with respect to the "true" values, the method could give different results about the number of selected parameters with a higher frequency.

The estimability method proposed by Wu et al., (2011) proved to be an efficient tool for overcoming estimability issues, producing results regarding parameters rank of influence that were coherent with other estimability methods used in this study. Moreover, it offers the distinct advantage of identifying which parameters should be selected for estimation, without the need of an arbitrary threshold or cut-off value, while maintaining low computational cost.

Results from the analysis of parameter estimates precision suggest that repeating the estimability method, using the parameter estimates from the previous iteration as initial guesses in subsequent iterations, can lead to more accurate rankings of parameter influence and a more accurate selection of parameters to be estimated, while also improving the precision of the estimates.

Future work should focus on applying the estimability method to more complex models with a larger number of parameters, as well as conducting a more comprehensive analysis of the interaction between model-based design of experiments and the estimability method. It would also be valuable to select model parameters based on the mean squared error of predictions at the operating conditions of interest, rather than limiting the analysis to experimental points. This approach, already explored in the literature, could lead to improved model performances, especially when operating conditions differ significantly from the experimental setup.

References

Beck, J. V. and K. J. Arnold (1977). *Parameter Estimation in Engineering and Science*. John Wiley & Sons, NY (U.S.A.), pp. 156-172.

Braakman, S., P. Pathmanathan and H. Moore (2022). Evaluation framework for systems models. *Pharmacometrics and Systems Pharmacology*, **11**, pp. 264-289.

Chu, Y. and J. Hahn (2009). Parameter Set Selection via Clustering of Parameters into Pairwise Indistinguishable Groups of Parameters. *Industrial & Engineering Chemistry Research*, **48**, pp. 6000-6009.

Chu, Y., Z. Huang and J. Hahn (2009). Improving Prediction Capabilities of Complex Dynamic Models via Parameter Selection and Estimation. *Chemical Engineering Science*, **64**, pp. 4178-4185.

Degenring, D., C. Froemel, G. Dikta and R. Takors (2004). Sensitivity Analysis for the Reduction of Complex Metabolism Models. *Journal of Process Control*, **14**, pp. 729-745.

Foss, B.A., B. Lohmann and W. Marquardt (1998). A field study of the industrial modeling process. *Journal of Process Control*, **8**, pp. 325-338.

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

Galvanin, F., M. Barolo and F. Bezzo (2009). Online model-based redesign of experiments for parameter estimation in dynamic systems. *Industrial and Engineering Chemistry Research*, **48**, pp. 4415–4427.

Galvanin, F., S. Macchietto and F. Bezzo (2007). Model-based design of parallel experiments. *Industrial and Engineering Chemistry Research*, **46**, pp. 871–882.

Grewal, M. S. and K. Glover (1976). Identifiability of Linear and Nonlinear Dynamical Systems. *IEEE Transactions on Automatic Control*, **21**, pp. 833–836.

Holmberg, A. (1982). On the Practical Identifiability of Microbial Growth Models Incorporating Michaelis-Menten Type Nonlinearities. *Mathematical Biosciences*, **62**, pp. 23-43.

Jiménez-Hornero, J. E., I. M. Santos-Dueñas and I. García-García (2008). Structural Identifiability of a Model for the Acetic Acid Fermentation Process. *Mathematical Biosciences*, **216**, pp. 154-162.

Kim, B. and J. H. Lee (2019). Parameter subset selection and biased estimation for a class of ill-conditioned estimation problems. *Journal of Process Control*, **81**, pp. 65-75.

Kubokawa, T., C. P. Robert, A. K. Md and E. Saleh (1993). Estimation of Noncentrality Parameters. *Canadian Journal of Statistics*, **21**, pp. 45-57.

Lam, N. N., P. D. Docherty and R. Murray (2022). Practical identifiability of parametrised models: A review of benefits and limitations of various approaches. *Mathematics and Computers in Simulation*, **199**, 202-216.

Leis, J. R. and M. A. Kramer (1988). The Simultaneous Solution and Sensitivity Analysis of Systems Described by Ordinary Differential Equations. *ACM Transactions on Mathematical Software*, **14**, pp. 45-60.

Ljung, L. and T. Glad (1994). On Global Identifiability for Arbitrary Model Parametrizations. *Automatica*, **30**, pp. 265-276.

Mclean, K. A. P. and K. B. Mcauley (2012). Mathematical modelling of chemical processes-obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *Canadian Journal of Chemical Engineering*, **90**, pp. 351-366.

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

Miao, H., X. Xia, A.S. Perelson and H. Wu (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *Society for Industrial and Applied Mathematics*, **53**, pp. 3-39.

Petersen, B., K. Gernaey and P. A. Vanrolleghem (2001). Practical Identifiability of Model Parameters by Combined Respirometric-Titrimetric Measurements. *Water Science and Technology*, **43**, pp. 347-355.

Pohjanpalo, H. (1978). System Identifiability Based on Power Series Expansion of Solution. *Mathematical Biosciences*, **41**, pp. 21-33.

Quaglio, M. (2020). Novel techniques for kinetic model identification and improvement. PhD thesis, University College London.

Raue, A., C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller and J. Timmer (2009). Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood. *Bioinformatics*, **25**, pp. 1923-1929.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*. 2nd ed., Duxbury Press, Belmont, CA (U.S.A.), pp. 175-180.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, **145**, pp. 280-297.

Saltelli, A., K. Chan and E. M. Scott (2000). *Sensitivity Analysis*. John Wiley & Sons, Inc., NY, pp. 56-61.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola (2008). *Global Sensitivity Analysis. The Primer*. John Wiley and Sons Ltd, West Sussex (England), pp. 11-33.

Schittkowski, K. (1986). NLPQL: A fortran subroutine solving constrained nonlinear programming problems. *Ann Oper Res*, **5**, pp. 485-500.

Sobol, I.M. (1993). Sensitivity estimates for nonlinear mathematical models, *MMCE*, **1**, pp. 407-414.

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

Vajda, S. and H. Rabitz (1989). State Isomorphism Approach to Global Identifiability of Nonlinear Systems. *IEEE Transactions on Automatic Control*, **34**, pp. 220-223.

Vajda, S., H. Rabitz, E. Walter and Y. Lecourtier (1989). Qualitative and Quantitative Identifiability Analysis of Nonlinear Chemical Kinetic-Models. *Chemical Engineering Communications*, **83**, pp. 191-219.

Wu, S., K. A. P. McLean, T. J. Harris and K. B. McAuley (2011). Selection of optimal parameter set using estimability analysis and MSE-based model-selection criterion. *International Journal of Advanced Mechatronic Systems*, **3**, pp. 188-197.

Wu, S., T. J. Harris and K. B. McAuley (2007). The Use of Simplified or Misspecified Models: Linear Case. *The Canadian Journal of Chemical Engineering*, **85**, pp. 386-398.

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