Identification of Dynamic Metabolic Flux Balance Models Based on Parametric Sensitivity Analysis

by

Ricardo Martinez Villegas

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Chemical Engineering

Waterloo, Ontario, Canada, 2016

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

A dynamic mathematical model that involves a set of physicochemical parameters can describe a dynamic system. Parametric sensitivity analysis studies the effect of changes in these parameters on model outputs of interest. If a system is operated within a region of high sensitivity any small change in the parameter values drastically affect the output. Hence, it is essential to be able to predict this sensitivity when designing, operating or optimizing a system based on the model. Dynamical biological models that describe gene regulation, signalling, and metabolic networks are strongly dependent on a large number of parameters. Most of these models are highly nonlinear and involve a high-dimensional state space. Conventional parametric sensitivity analysis that examines the effect of each parameter independently at one specific moment is generally inaccurate since it ignores correlations between parameters. Thus, it is very important to account for correlations when conducting a parametric sensitivity analysis.

Model parameters are never known accurately and consequently they are typically described by a range of values. Some parameters may be measured directly but even for such case they will exhibit variability due to noise, e.g. a flow rate that is measured by a noisy flow meter. The variability in values of model parameters that cannot be measured directly arises from two main sources: i- noise in data and ii- process disturbances that translate directly or indirectly into changes in the parameters. In the presence of measurement noise, the identification of model parameters from data will result in model parameter values that are known within bounds with different levels of confidence. Also, process disturbances may directly affect the value of a parameter, e.g. changes in initial conditions of a metabolite concentration in a batch culture, or indirectly, e.g. changes in oxygen transfer due to changes in aeration rates.

This thesis focuses on the identification of model parameters for biochemical systems. Models describing such systems are based on the biochemical reactions occurring within an organism that are used to produce or consume essential components to grow, reproduce, preserve cell

structures, and respond to environmental changes. This group of reactions is collectively referred to as a metabolic network.

Dynamic Flux Balance Analysis, a particular modeling method, which is the focus of the current work, can be used to study microbial metabolic networks. This type of mathematical model can simulate the metabolism of an individual cell by describing the flux distribution inside a cellular network. The approach is based on maximizing a biological based objective such as growth rate or production of ATP subject to constraints on the rate of change of certain metabolites. Several other approaches have been developed in turn to simulate the responses of the cells to different stimulus. Nevertheless, Dynamic Flux Balance Analysis compared to other approaches is advantageous in terms of the relatively smaller number of parameters that have to be calibrated to fit the data thus resulting in lower sensitivity to noise and requiring smaller data sets for calibration. In view of its advantages, this thesis focuses on this particular modeling approach, which is becoming increasingly popular in the field of biotechnology and systems biology disciplines.

The research to be presented will focus on the robust identification of dynamic metabolic flux models based on parametric sensitivity analysis. The particular case study that is chosen to illustrate the proposed method is Diauxic growth in *Escherichia coli* in a batch culture. This approach intends to show how to identify the model parameters of the dynamic model based on a parametric sensitivity analysis that explicitly accounts for correlations in the data. The sensitivity is quantified by a parameter sensitivity spectrum. Then, the parameters are ranked based on this analysis to assess whether a subset of the parameters can be eliminated from further analysis. Finally, identification of the remaining significant parameters is based on the maximization of an overall parametric sensitivity measure subject to set based constraints that are derived from the available data. The parametric sensitivity method is global in the sense that it examines the simultaneous variation of all the model's outputs instead of focusing on outputs variables one at a time.

Acknowledgments

I would like to thank first of all my supervisors Dr. Hector Budman and Dr. Ali Elkamel for their support, their patience, their advices, the encouragement and preparation they gave me, and the knowledge they shared with me. But most of all for the opportunity they gave me to come here to fulfil one of my biggest dreams, acquire a Master's degree.

I want to thank also CONACYT (Consejo Nacional de Ciencia y Tecnologia) for providing me with the financial support that I needed to complete my degree.

Also I would like to thank my colleagues and friends, Rubin Hille, Omar Santader, Vanessa Zavatti, and Yuncheng Du, for all the support they gave me and having patience with me specially when I was learning how to code in Matlab and helping me with the definitions of the complicated words that at the beginning I did not understand from my supervisors.

I am also very thankful with all the friends and colleagues from the University of Waterloo that encouraged me to work towards obtaining my degree. Guillermo Rodriguez, Ushinik Mukherjee, Pedro Cordero, Edoardo Lazzaroni, Nicolas Conda, Mohamed, and Rogelio Hernandez, thanks guys for sharing so many good times during these years.

And finally I want to thank all my friends and family in Mexico who also encouraged me to continue with my dreams.

Dedication

To my parents, Esther Villegas and Jose Martinez, my sister, Karina Martinez, and my girlfriend, Iulia Andrasi. Without their support, believe and encouragement I would not be here completing this dream.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 RESEARCH MOTIVATIONS

Any kind of chemical process involves a series of steps to transform one or more compounds into a desired compound. These series of steps conform all together a methodology for the process or a system. The transformation can occur spontaneously or may be driven by an external force, but in general it involves chemical reactions. Specifically, biochemical processes use organisms or biomolecules to perform all the chemical reactions that lead to the production of a specific biological compound. Chemical and biochemical processes are governed by different physicochemical factors, which affect directly or indirectly the behaviour of the system.

All these systems, and specially the biological ones, are highly complex because they exhibit nonlinear dynamic behaviour. This complexity makes them difficult to analyze and control. Gaining understanding about a bioprocess is essential for maximizing productivity and improving quality of the bio product. Mathematical models can be used to approximate the behaviour of dynamic bio-systems and for predicting their variability with respect to different physicochemical factors.

Parametric Sensitivity Analysis (PSA) studies the changes and the system's sensitivity with respect to a specific parameter set. By performing a PSA one will be able to identify the relative effect of the set of parameters on the system's outputs. In particular, one can identify the parameters that have very small influence on the system's outputs so as to simplify the parameter estimation procedure. Thus, parametric sensitivity can be used for both studying the sensitivity of the system with respect to changes in parameters and for simplifying the parameter estimation method.

1.2 DYNAMIC BIOLOGICAL MODELS

Biological Systems are highly complex systems that involve a set of biochemical reactions to transform nutrients into products.

Experimental studies of biological systems are generally time consuming, expensive and prone to contamination problems. Mathematical modelling of dynamic biological models is becoming a valuable technique to replicate and study mechanisms used by cells to reproduce and produce molecules of therapeutic or industrial use. Thus, mathematical modeling can be used to save expensive experimentation time and costs. By using computational techniques researchers have been able to observe and simulate simultaneously the behaviour of dynamic systems and analyze the production and consumption of distinct molecular species.

Two new disciplines had emerged in recent years that used mathematical modelling as their basic tool: Systems Biology and Synthetic Biology. Systems Biology promotes the use of computational tools to characterize the behaviour of biological networks while Synthetic Biology focuses on designing and building genetic networks based on genetic models' predictions. The tools developed in these two fields are being implemented in areas of health and disease treatment, bioprocess engineering, pharmaceutics and vaccine manufacturing, renewable energies, environmental remediation, etc.

The use of dynamic biological models helps quantifying the interactions occurring inside the cell. One example could be the binding or unbinding of two species, where we need to provide rates associated to either the separation of the molecules and/or the attachment of the molecules to each other. There are large databases available that provide information about binding and dissociation interactions between species for many types of compounds and organisms (Ingalls, 2013).

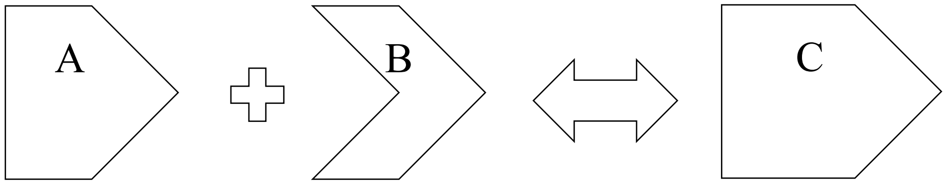

Figure 1.1 Specie A binds reversibly Specie B to create Specie C

1.3 SENSITIVITY ANALYSIS

Dynamic models are used to approximate and study the behaviour of complex systems. Sensitivity analysis of a model helps identifying which parameters have most effect on the outputs of the system, i.e. which ones contribute the most to the variability of the system's outputs and the correlation among parameters and how that correlation affects the outputs. The less significant parameters can be either eliminated to reduce the complexity of the model or kept at a constant arbitrary value without significant effect on the outputs (Iman and Helton, 1988).

Sensitivity analysis typically involves the following steps: -i define the model to be analyzed; ii- identify the dependent and independent variables in the model, iii- define an appropriate probability function for the input parameter, iv- generate a suitable sampling method for the output variables; and v- calculate and analyze the contribution of one or a group of input parameters on the set of outputs (Iman et al., 1981a; Iman et al., 1981b; Helton et al., 1985; Helton et al., 1986).

There are several techniques to perform sensitivity analysis of parameters affecting dynamic and static models. These techniques are generally classified into two main groups: Local Sensitivity Analysis, which purpose is to analyze the effect of each parameter on each output; and Global Sensitivity Analysis, which looks for the effect of simultaneous variations of parameters. The advantage of global sensitivity analysis is that it takes into account the effect of correlations among parameters.

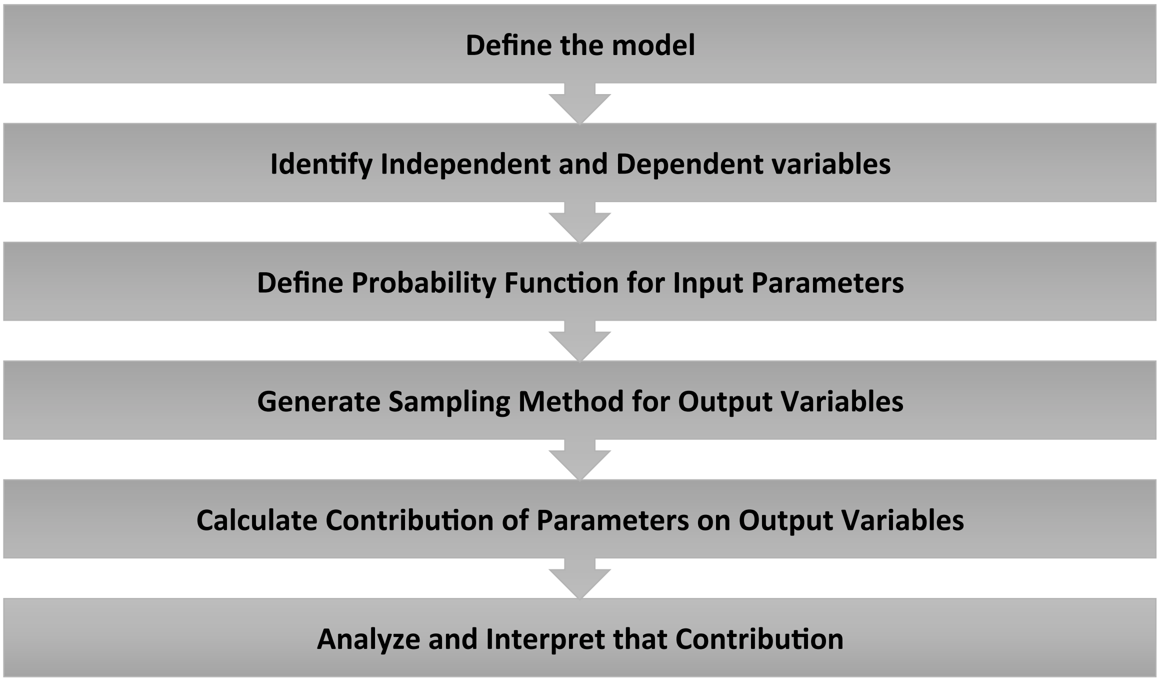
Define the model
Identify Independent and Dependent variables
Define Probability Function for Input Parameters
Generate Sampling Method for Output Variables
Calculate Contribution of Parameters on Output Variables
Analyze and Interpret that Contribution

Figure 1.2 Diagram of the steps to perform a Sensitivity Analysis of a model

1.4 PARAMETER ESTIMATION

Parameter estimation is the procedure by which parameter values are calculated by matching the values calculated from the model predictions to the corresponding data. The estimation procedure typically assumes some knowledge about the statistical distributions of the data and the parameters to be estimated (Michiels et al., 2002). The mathematical approach used to calculate the values from the parameters is referred to as the estimator, and its result is defined as the estimate. The accuracy of the approximation depends on the value of the standard deviation of the estimate (van den Bos, 2007).

Parameter estimation procedures are commonly used to both to calibrate and validate mathematical models with data. Most of the observable values contain variability due to noise and disturbances. Experiments always differ from each other even for cases that they are conducted at identical measurable operating conditions. This variability is generally characterized using statistical methods (Bard, 1974).

There are two main approaches in parametric estimation that depend on the type of model. Linear estimation or nonlinear estimation techniques can be chosen for estimating parameters depending on the level of nonlinearity of the process under study (Englezos and Kalogerakis, 2001).

1.5 RESEARCH OBJECTIVES

The main objective of this research is to estimate the number of model parameters and their values based on the sensitivity analysis of the outputs of a dynamic biological model. A Diauxic growth model of *E. coli* is used as the case study. A parameter estimation procedure with a global sensitivity analysis approach is used for this research.

The purpose of the current research is then summarized as follows:

1. Analyze and understand the behaviour of a Dynamic Biological Model.
2. Identify the parameter set and the correlation of parameters that highly affect the model's outputs.
3. Perform model calibration using information from the Sensitivity Analysis.

1.6 OVERVIEW

This thesis includes five chapters. Chapter 2 presents the theoretical framework about kinetic modelling, metabolic network modelling, parametric sensitivity analysis, and parameter estimation. Chapter 3 covers the methodology implemented in this research and the case study that is sued to illustrate the proposed methodologies. In Chapter 4 the results obtained from the analysis performed in chapter 3 are presented and analyzed. Chapter 5 contains the final conclusions and recommendations for future work.

Chapter 2

LITERATURE REVIEW

Chapter Outline This chapter presents a summary of the most important concepts used in this thesis.

2.1 KINETIC MODELING

Modelling of chemical and biochemical reactions helps engineers to visualize the transformation of compounds involved in these reactions and the interactions between the compounds while they are being converted. A model can also predict production and consumption rates thus making it a very useful tool when trying to perform optimization.

To formulate a kinetic model for a bioprocess it is necessary to identify the reaction pathways for a particular molecule when being transformed into different molecules, and the dynamic mass balances described by differential equations based on equilibration of production and consumption rates (Rahul, 2012).

The steps involved in the formulation of a dynamic kinetic model are:

1. Identify the chemical or biochemical pathways,
2. Assume kinetic expressions and calculate mass balances using these expressions,
3. Calculate the best numerical values of the kinetic parameters necessary to fit the available experimental data.

The information for step 1 can be obtained from experimental data or from databases like KEGG, Brenda and EMP.

2.1.1 Chemical Reaction Networks

The transformation of one or more chemical reactants into products involves a set of reactions. This set constitutes a network that has to be identified in order to formulate a model. This network of reactions can be typically represented in graphical form to facilitate understanding of the process For example; a set of reactions conforming a network is represented by:

$$A + B = C$$

$$C = D + E$$

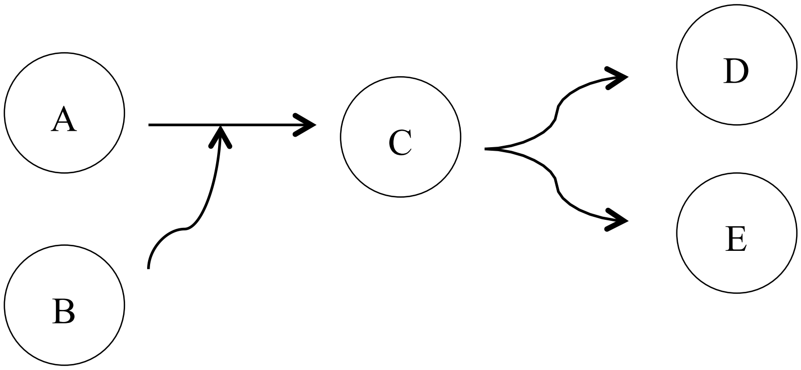

Figure 2.1 Exemplification of a reaction network.

Where, Figure 2.1 is a graphical representation of the chemical network.

There are two kinds of chemical reaction networks: closed and open networks. Closed networks are groups of reactions where the reactants and the products remain within the network. When closed networks reach equilibrium or steady state, the reaction rates are zero. Contrarily, open networks are groups of reactions that exchange components within and outside the network. When open networks reach steady state the system is considered to be in a dynamic equilibrium and the inputs' rates of consumption are equal to the rate of outputs' production.

Biochemical processes are typically described by open networks because in such processes nutrients are externally supplied, e.g. through feeding of growth media into the system and secreting metabolites or wastes outside of the system (Ingalls, 2013).

Another important element to consider when building a biochemical model is the directions of the reactions. Chemical reactions can be reversible or irreversible. In reversible reactions reactant or reactants will either form a product or they will be formed from that product. On the other hand irreversible reactions can only occur in one direction, e.g. certain nutrients can only be consumed but cannot be produced.

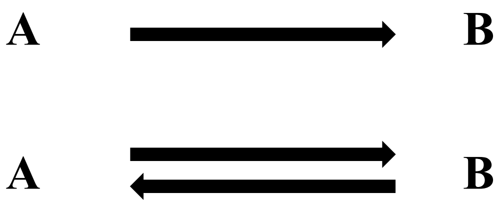

Figure 2.2. Examples of irreversible and reversible reactions

Finally, reaction rates have to be considered and identified for properly describing the dynamic evolution of metabolites. The reaction rates are the measure of how fast one or more reactants will be transformed into a specific product. The reaction rate or rate of change of one species over time can be modeled using an Ordinary Differential Equation (ODE). The ODE is formulated by equating the difference between the rates of production minus the rate of consumption to the rate of change from each of the compounds in the system.

$$\underbrace{\frac{d}{dt}[A]}_{\text{Rate of Change}} = \underbrace{k2[A]}_{\text{Production Rate}} - \underbrace{k1[A]}_{\text{Consumption Rate}}$$

Figure 2.3. Reaction Rate model

Since the last equation will typically contain nonlinear terms due to the nonlinearity of reaction rate expressions numerical methods will be used for finding approximate solutions. Using these methods the only inputs needed are the initial values and the time scale for the reaction rates.

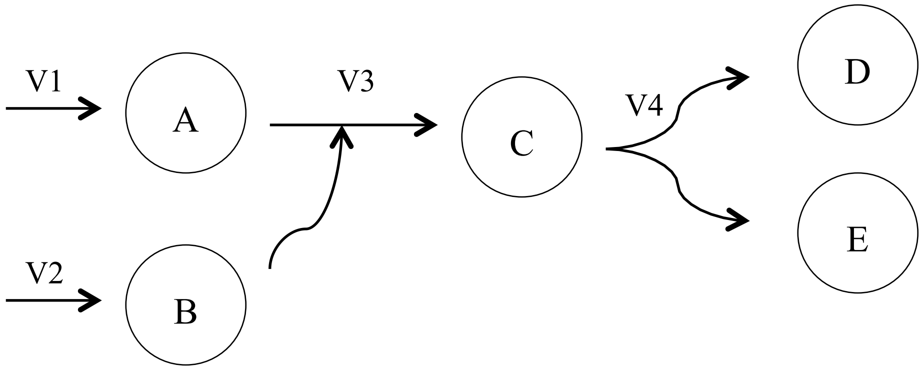

Figure 2.4. Reaction network with reaction rates labelled V_i

To mathematically model a network with all its possible reactions we need to define all the reaction rates, represented by V_i in Figure 2.4, that determine the production and consumption rates. The differential equations that are used to balance each metabolite in a network will be constructed using the production and consumption rates for each of the species using a plus sign when the metabolite is being produced and a minus sign when it is consumed. The set of all the differential equations can be then simulated for different scenarios such as different nutrient rates or different initial conditions to gain understanding about the system's behaviour.

Example of a Reaction Network

Defining Vi for each rate and using figure 2.4: V1=k1, V2=k2, V3=k3[A][B], V4=k4[C]

$$\frac{d}{dt}[A] = k1 - k3[A][B]$$

$$\frac{d}{dt}[B] = k2 - k3[A][B]$$

$$\frac{d}{dt}[C] = k3[A][B] - k4[C]$$

$$\frac{d}{dt}[D] = k4[C]$$

$$\frac{d}{dt}[E] = k4[C]$$

(2.1)

2.1.2 Biochemical Networks

All biological reactions rely on the action of enzymes. Enzymes are proteins that have the role of regulating most of the internal and external processes from cells. The enzymes bind their substrates in a lock-key arrangement, due to the specificity of their structure, and catalyze the transformation of these substrates into a more useful compound for the cell. The catalytic processes regulated by the enzymes can be classified into two main groups: anabolic and catabolic reactions. Anabolism is the binding of different substrates into a bigger compound, whereas, Catabolism is the decomposition of big structures into simpler ones. Both activities are essential for the cell and the coupling of these two reaction mechanisms is essential for explaining metabolic phenomena.

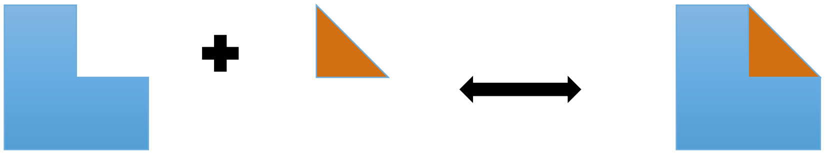

Figure 2.5. Enzyme-substrate binding model

Enzymes are conformed by long chains of amino acids. Due to the physicochemical properties of these chains they can fold in different ways. Folding results in the creation of open spaces within the enzyme (blue shape in Figure 2.5), which resemble voids. The boundaries of these voids exhibit electrochemical activity because of the amino acids present at these boundaries. These voids are referred to as active sites. The active sites allow molecules called substrates (orange in Figure 2.5) to bind with the amino acids in the enzyme thus resulting in conformational changes in the structure of the substrate. The shape of the active site has a specific form and only allows specific substrates with the same dimensions to bind and interact with the amino acids in the enzyme. This specificity of enzymes towards a given substrate makes the reactions that occur inside a cell to be very efficient. The direction of these reactions depends on the concentration gradients (Figure 2.5).

In 1913 Leonor Michaelis and Maud Menten created a mathematical model to describe the dynamics of the enzyme-substrate binding. This kinetic model has helped researchers to understand and study the catalytic reaction mechanism used by the enzymes. The Michaelis-Menten model hypothesizes that the available enzyme in the media binds the substrate creating a complex, which is transformed into a product and a remaining quantity of free enzyme. The assumption behind the development of the kinetic rate expression is that the product never binds back with the enzyme and the reaction is almost instantaneous (Rahul, 2012). On the other hand the intermediate step of binding of the enzyme with the substrate described in Figure 2.5 is assumed to be reversible.

$$[E] + [S] \leftrightarrow_{k-1}^{k1} [ES] \rightarrow^{k2} [P] + [E]$$

Figure 2.6. Michaelis-Menten kinetic model

Applying the law of mass action to the model in figure 2.6 the following system of differential equations results:

$$\frac{d}{dt}[E] = k_{-1}[ES] - k_1[E][S] + k_2[ES]$$

$$\frac{d}{dt}[S] = k_{-1}[ES] - k_1[E][S]$$

$$\frac{d}{dt}[ES] = -k_{-1}[ES] + k_1[E][S] - k_2[ES]$$

$$\frac{d}{dt}[P] = k_2[ES]$$

(2.2)

Since it is really difficult to measure the amount of free enzyme and the concentration of enzyme-substrate compound over time the following equation is used to describe the concentration of initial enzyme, which is assumed to be known a priori, $[E] + [ES] = [E_0]$. This relation can be used for solving the free enzyme concentration over time in the mass balance equations. Applying a fast equilibrium assumption whereby the substrate is immediately converted to product, solving for the concentration of enzyme and substrate compound and substituting the resulting equation in the substrate-enzyme differential equation balance we obtain (Ingalls, 2013):

$$[E] = [E_0] - [ES]$$

$$\frac{d}{dt}[S] = k_{-1}[ES] - k_1([E_0] - [ES])[S] = 0, \qquad [ES] = \frac{[E_0][S]}{\frac{k_1}{k_{-1}} + [S]}$$

$$\frac{d}{dt}[P] = \frac{k_2[E_0][S]}{\frac{k_1}{k_{-1}} + [S]}$$

(2.3)

Renaming $v = \frac{d}{dt}[P]$, $Vm = k_2[E_0]$, and $Km = \frac{k_1}{k_{-1}}$ it is possible to obtain the Michaelis-Menten expression describing the rate of formation of the product as follows:

$$v = \frac{Vm * [S]}{Km + [S]}$$

(2.4)

The equation above is a good approximation to describe how a microorganism consumes a substrate and produces one specific metabolite, but it cannot be used for all cases because many enzymes do not behave in this way. For example, some product formation reactions are reversible, some reactions involve two or more substrates inside the enzyme, some compounds inhibit enzyme activity and the activity of enzymes may vary according to specific regulatory mechanisms, e.g. allosteric or cooperatives regulations. Additional information on mathematical modeling of biological systems can be found elsewhere (e.g. Ingalls, 2013)

2.2 MODELLING METABOLIC NETWORKS

Mathematical modeling is an attractive alternative to experimental techniques due to the relatively lower associated costs as compared to experiments. Modelling metabolic networks can be, however, a complicated task, mainly because of the vast array of regulatory actions and complexity occurring in biological systems. The amount of detail put in the development of dynamic metabolic network model will determine its accuracy and reliability. Models that are based on the metabolic network can predict the behaviour of a dynamic biological system well.

Different metabolic networks model have been developed depending on the needs of researchers. Examples of these methods are Flux Balance Analysis, Elementary Flux Modes Analysis and Metabolic Flux Analysis. The intention of these models is to predict the flux distribution occurring inside the cell either at steady state or transient state. Metabolic models use stoichiometric information to calculate fluxes that are responsible for the production of a certain compound of interest. The requirements of computational data to adjust the model and the time required to fit the model are the criteria that makes one method more attractive than another (Wang, 2011).

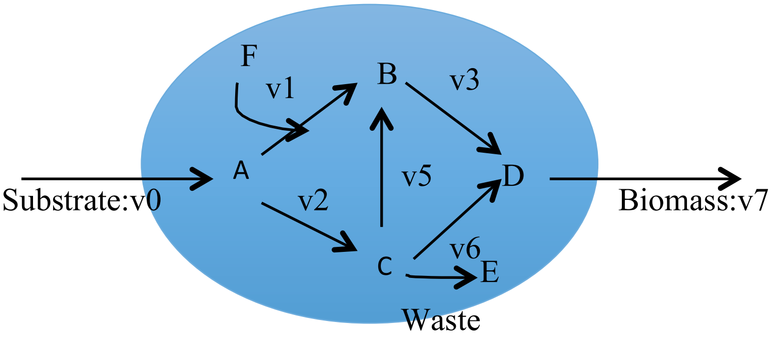

Figure 2.7 Simple Microbial Network

Flux Balance Analysis (FBA) is a constrained steady-state optimization method that can predict the consumption and production rates of metabolites in the metabolic network of a microorganism and can predict the distribution of fluxes inside the cell of larger biological systems. A flux is the amount of consumed or produced metabolite per unit time and per unit cell associated with a particular reaction involving that metabolite. The methodology works by optimizing a certain objective function, which in most of cases is the Biomass growth, subject to stoichiometric constraints, thermodynamic constraints that determine the direction of the reactions and maximum uptake rate constraints (Sharan, 2006). Other objectives have been considered such as maximization of ATP productivity or maximization of substrate consumption per unit flux (Varma and Palsson, 1994a; Orth et al., 2010).

FBA has become in the last decade one of the most utilized techniques to approximate the functioning of metabolic networks. In general it requires less input data since the cell behaviour is typically determined by a maximization of an objective subject to a limited number of constraints whereas the metabolites that are not constrained can be derived from stoichiometry. The fact that it requires less input data and no kinetic information makes it a good alternative to experimental approaches (Varma and Palsson, 1994a). Flux Balance Analysis has given promising steady state predictions and has been shown to approximate experimental results well (Kauffman et al., 2003).

The procedure to perform a Flux Balance Analysis is as follows:

1. Define and construct the metabolic network of the organism that includes all reactions and all the metabolites found in the network. Public databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) can be used to define the network (Kanehisa et al., 2010).

2. Create a mathematical representation of the stoichiometry of the metabolic reactions, matrix (***S***), with all the stoichiometric coefficients corresponding to each reaction occurring in the network. This is an *m x n* matrix, where every row (*m*) represents one compound and every column (*n*) represents one reaction (Figure 2.8). The values entered

in the columns represent the coefficients of the metabolites from each reaction. Negative values are used for consumption and positive values for production of metabolites (Orth et al., 2010).

Matrix **S**	Reactions
Metabolites	$\begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & \dots & & A_{2,n} \\ \vdots & & \dots & \vdots \\ A_{m,1} & A_{m,2} & \dots & A_{m,n} \end{bmatrix}$

Figure 2.8. Stoichiometric Matrix.

3. Define a flux for each reaction in the network. Fluxes are represented with the letter v, and have a dimension of $n\ x\ 1$**.** The Flux Balance Analysis determines the fluxes inside the network and is based on the following equation: $\boldsymbol{S} * \boldsymbol{v} = \boldsymbol{b}$**.**

4. The next step is determining the objective function to be maximized or minimized. This function is typically a linear combination of some fluxes: $\boldsymbol{Z} = \Sigma\, \mathbf{w}_{ij}\, \boldsymbol{v}_i = \boldsymbol{w}^T \boldsymbol{v}$, where $\boldsymbol{w}$ is the weight contribution vector for each flux. The most common objective function is the maximization of the growth rate. The optimization function is subject to constraints in the rate of change of metabolites within upper and lower bounds (a_i and b_j) (Hjersted and Henson, 2006; Mahadevan et al., 2002; Orth et al., 2010).

$$\max \boldsymbol{Z} = \boldsymbol{w}^T \boldsymbol{v}$$

or

$$\min \boldsymbol{Z} = \boldsymbol{w}^T \boldsymbol{v}$$

Subject to:

$$\boldsymbol{S} * \boldsymbol{v} = \boldsymbol{b}$$

$$a_i \leq v_j \leq b_j \tag{2.5}$$

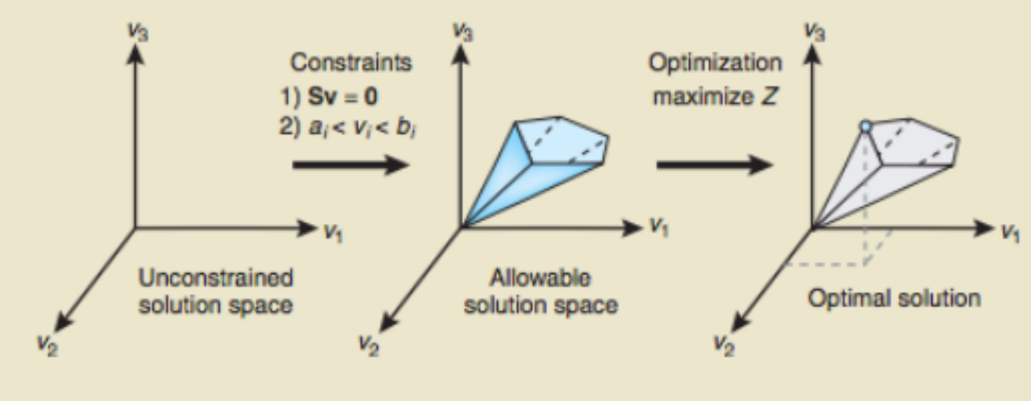

Figure 2.9. FBA Modeling with Constraints obtained from (Orth et al., 2010).

The disadvantage of classical FBA is that it does not take into account dynamic behaviour. For example, Mahadevan et al showed for the diauxic growth in *Escherichia coli* that the FBA incorrectly estimates the time of re-utilization of acetate following glucose depletion (Mahadevan et al., 2002; Niklas, et al., 2010). Regular Flux Balance Analysis cannot estimate the production or consumption of metabolites over time. Moreover, the genetic regulations of the reactions and kinetic rates are not considered in view that FBA is based on the assumption of steady state. For these reasons, an extension of the Flux Balance Analysis has been suggested that is referred to as Dynamic Flux Balance Analysis (DFBA) (Mahadevan et al., 2002; Hjersted et al., 2007; Nowruzi et al., 2008; Budman et al., 2013) which is the focus of the current thesis.

Dynamic Flux Balance Analysis can model the dynamics of biological systems by imposing rate of change constraints on fluxes at each time interval (Mahadevan et al., 2002). These rate constraints are typically expressed with kinetic expressions such as Monod equation as a function of the time varying concentrations of substrates and products involved in the reaction associated with the constrained flux. The dynamic flux model offers a practical approach to develop a full metabolic network model when not enough kinetic data is available (Hjersted and Henson, 2006). This methodology is based on the assumption that the concentrations of metabolites equilibrate fast in response to disturbances. The kinetic data from the substrate uptake and production rates can be incorporated by coupling the dynamic mass balances with the stoichiometric model (Stephanopoulos et al., 1998). The metabolites concentrations are typically evolved with respect to time according to the equation:

$$\frac{d\boldsymbol{Z}}{dt} = \boldsymbol{S} * \boldsymbol{v} * X \tag{2.6}$$

Where Z is the vector of metabolites' concentrations, S is the stoichiometric matrix, $\boldsymbol{v}$ is the vector of fluxes and X is the biomass concentration.

2.3 PARAMETRIC SENSITIVITY ANALYSIS

Varma et al. (1999) defines the parameters as the physicochemical factors that alter the functioning of any type of system. However, since the current thesis focuses on model identification, we define the parameters as mathematical variables that have to be calibrated to provide good match between model predictions and data. The analysis of the effect of these parameters on the model outputs is called parametric sensitivity.

Parametric sensitivity analysis may serve for a number of tasks as follows: (Iman and Helton, 1988; Hamby, 1995):

i. Model reduction where certain parameters are eliminated or fixed at a nominal value since they do not affect significantly the output.
ii. Limiting the calibration of the model with respect to the parameters with the highest influence on outputs
iii. Find the correlations between the parameters involved in the model; and
iv. Identify the parametric sensitivity region where the parameters have the bigger effect on the model outputs.

Several techniques are available (Figure 2.10) to perform sensitivity analysis. These techniques can be generally classified into two main groups: local sensitivity analysis and global sensitivity analysis. Local sensitivity studies consider independent variations around a determined parameter set. Global sensitivity studies take into account larger and simultaneous variations in the parameters with the purpose of finding correlations between them (Varma et al., 1999).

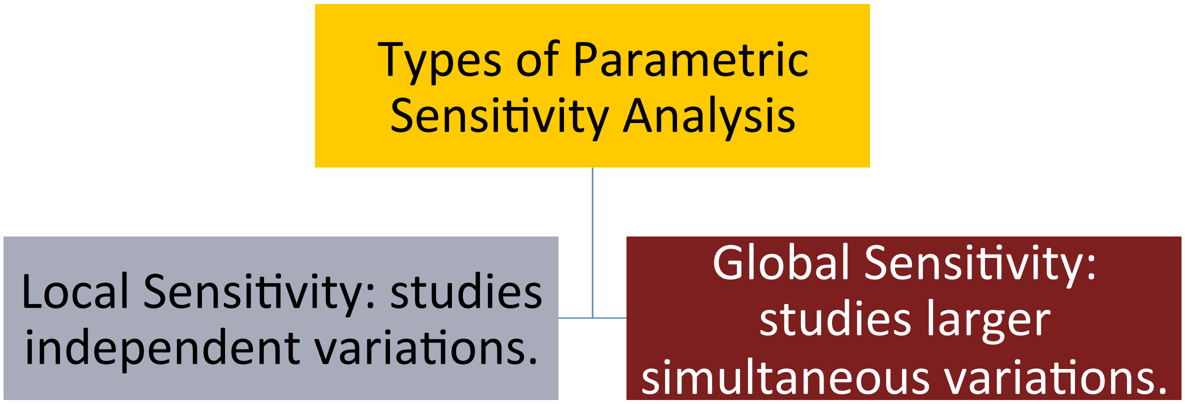
Types of Parametric Sensitivity Analysis
Local Sensitivity: studies independent variations.
Global Sensitivity: studies larger simultaneous variations.

Figure 2.10. Parametric sensitivity approaches.

2.3.1 Local Sensitivity

Local sensitivity analysis studies the effect that individual parameters have on the model outputs in a given region of space (i.e. around a nominal value). The analysis involve generating small changes in the nominal values of the input parameters one at a time, and then analyze the effect of those changes on the output or dependent variables. This analysis can be quantified using partial differentiation involving the output variables over time and the input parameters.

Let us define the output variable *y* that depends on time (*t*) and the input parameter (*x*). This will result in the following relation: $y = y(t, x)$. To perform a local sensitivity analysis, a small change in *x* is generated with respect to its nominal value and the effect that this change has on *y*, $y = y(t, x + \Delta x)$ is calculated. In the limit when $\Delta x \rightarrow 0$ the changes in *y* with respect to changes in *x* can be expressed by the following partial derivative:

$$s(y; x) = \lim_{\Delta x \rightarrow 0} \frac{y(t, x + \Delta x) - y(t, x)}{\Delta x} = \frac{\partial y(t, x)}{\partial x}$$

(2.7)

The equation above is referred to as a first order local sensitivity of the dependent variable (*y*) with respect to the input parameter (*x*) (Varma et al., 1999). It is also possible to expand the local sensitivity analysis by defining local sensitivities based on higher-order expansions of the output y with respect to the input x. However most of the local sensitivity applications use the first order sensitivity given by equation 2.7.

Mathematical models of complex processes involve many parameters and outputs with different magnitudes and different units. For these reasons, when performing sensitivity analysis it is important to normalize all variables used in the analysis with respect to their nominal values.

$$S(y; x) = \left(\frac{x}{y}\right)_n * \frac{\partial y}{\partial x} = \left(\frac{x}{y}\right)_n * s(y; x)$$

(2.8)

Different types of computational techniques are available to perform local sensitivity of systems. Table 2.1 summarizes the three most common local sensitivity computational methods (Varma et al. (1999)).

2.3.2 Global Sensitivity

Local sensitivity analysis is sometimes unreliable because it ignores correlations and interactions among parameters. Global sensitivity analysis focuses on studying larger and simultaneous variations of a subset or all the parameters over the dependent or output variables of the model thus accounting for correlations among parameters. Global sensitivity analysis can also be used to quantify the uncertainty in parameter values in the presence of correlations (Cacuci et al., 2003; Saltelli et al., 2004; Campolongo et al., 2007; Saltelli et al., 2008).

Correlations among parameters are especially pervasive in biochemical models such as the ones used in this thesis due to the extensive use of Michaelis-Menten kinetics (equation 2.4). In equation 2.4 the parameters in the numerator and denominator are highly correlated. For instance if the substrate concentration is small, the data will be only informative about the ratio between the numerator and denominator parameters and thus it is irrelevant to test the effect of independent effect of numerator and denominator parameters on the output as done in local sensitivity analysis. Assessing the independent effect of numerator and denominator parameters on the outputs may lead to wrong and potentially too conservative or optimistic predictions of sensitivity. Thus, local sensitivity analysis cannot provide reliable information about the effect of simultaneous changes in parameters on the model outputs (Kitano, 2002; Kitano, 2004a; Stelling et al., 2004b). The robustness of biological models has attracted the attention of many researchers due to the complexity that these types of mathematical models exhibit (Kitano and Oda, 2006; Kitano, 2007a). Understanding the behaviour of these models in the presence of uncertainty in model parameters and being able to control them is also of paramount importance. Thus, global parametric sensitivity analysis of dynamic biological models is a more suitable tool to study robustness as compared to local sensitivity techniques.

Table 2.1. Local Sensitivity Computational Methods (Varma et al., 1999)

LSA Method	Algorithm description	Applications	Disadvantages
Direct Differential Method	The model and the sensitivity equations are solved simultaneously	- The number of dependent variables is smaller than the number of input parameters. - Sensitivities of output variables with respect to only few input parameters.	Stiffness problems might be encountered.
Finite Difference Method	Use of finite difference approximation to solve model equations and evaluate local sensitivities.	- The number of dependent variables is large. - Solving the model and sensitivity equation is not tractable. - Implicit objective for sensitivity analysis.	Finding a proper variation for each input parameter.
Green's Function Method	Solving first the homogenous part, and then with the use of linear integral transformations compute the local sensitivities.	- The number of dependent variables is much larger than the number of the input parameters. - Complete sensitivity analysis of all dependent variables.	Stiffness problems might be encountered.

Several approaches involving a series of numerical calculations have been developed to address global sensitivity analysis. All of these approaches assume random variability of the parameters with a probability and a cumulative density function for each parameter. In this way, it is possible to determine which parameters produce the maximum variance in the model outputs. In general, global sensitivity analysis proceeds as follows: i- assign a probability density function to every parameter; ii- generate samples using a sampling method within the parameter space; iii- calculate the outputs of the model on each sample point; and iv- quantify the sensitivity of the model based on a specific metric (Rahul, 2012). Global sensitivity methods are based on two main approaches, a Regression based method and a Variance based method. These methods are further discussed below and a graphical representation is shown in figure 2.11.

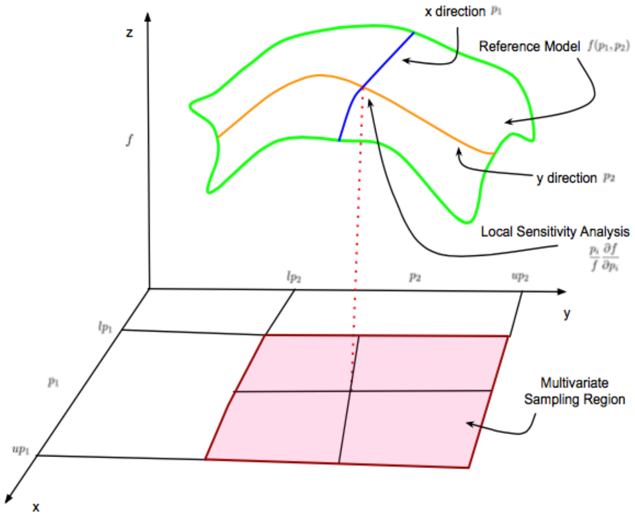
x direction p_1
z
Reference Model $f(p_1, p_2)$
f
y direction p_2
Local Sensitivity Analysis
$\frac{p_i}{f}\frac{\partial f}{\partial p_i}$
lp_2
p_2
up_2
y
lp_1
p_1
up_1
Multivariate Sampling Region
x

Figure 2.11. Graphic representation of a Global Sensitivity Analysis obtained from Rahul (2012)

Partial Rank Correlation Coefficient (PRCC) (Draper and Smith, 1998)

This is a regression-based method that uses a stratified sampling approach like the Latin hyper-cube to perform the analysis. This method is suitable for studying nonlinear models where a monotonic relation between inputs and outputs is found. To perform this analysis first the input data is ranked according to some criteria in an increasing order and the corresponding outputs are re-arranged accordingly. Then, a regression analysis is done on the ranked information and the Pearson rank coefficients are obtained (Blower and Dowlatabadi, 1994). This analysis determines how strong a correlation between the input and the output is. An important key assumption in this method is that the input variables are independent form each other.

Fourier Amplitude Sensitivity Test (FAST) (Cukier et al., 1973, 1975, 1978; Schaibly and Shuler, 1973; Koda et al., 1979; Mc Rae et al., 1982)

FAST is a variance-based method that finds the mean and variance values of the output variables and with these values it calculates the contribution that the inputs have on the output variances. The FAST method distributes the output's variances among the inputs and can be used to fix the parameters with no influence to their nominal value.

The methodology is based on the first-order parameter sensitivity calculated by the following equation (Cukier et al., 1978; Saltelli and Bolado, 1998):

$$S_i = \frac{D_i}{Var(Y)} \tag{2.9}$$

where the $Var(Y)$ represents the total variance of the output that is decomposed into increasing dimensionality terms (Cukier et al., 1978; Saltelli and Bolado, 1998):

$$Var(Y) = \sum_{i=1}^{n} D_i(Y) + \sum_{1 \le i \le j \le n}^{n} \left[D_{ij}(Y) + \cdots + D_{1\,2\ldots n}(Y) \right] \tag{2.10}$$

The terms D_{ij} are calculated based on Monte Carlo Sampling, and these elements contribute jointly to the variance of the model outputs. However those terms are difficult to calculate since they require extensive stochastic sampling from the parameter distributions, which is highly demanding for problems with many parameters. Consequently, this method was not extensively used until recently due to its computational complexity.

Extended Fourier Amplitude Sensitivity Test (eFAST) (Saltelli et al., 1992)

This method is an extension of the traditional FAST model. In FAST and eFAST the frequency response of model outputs is calculated with respect to model parameters by using Fourier Series representations for both parameters and outputs. The advantage of this method is that it calculates the total sensitivities and does not require first order approximations. FAST is only able to calculate the first order sensitivities, but by using eFAST we can obtain the total sensitivity measures, which is an estimation of the sum all the contributions from all the inputs as given by equation (2.11) (Saltelli and Bolado, 1998; Rahul, 2012). The numerator in 2.12 is related to the coefficients of the Fourier Expansions of the Output variables.

The whole contribution of the element X_i on the output is calculated with the addition of the first-order effect and all the rest high-order effects. In the scenario of a two-parameter model, the effect of the first element will be calculated as following:

$$D_1^{tot} = D_1 + D_{12} \tag{2.11}$$

and the total sensitivity will be given as:

$$S_{t1} = \frac{D_1^{tot}}{Var(Y)} \tag{2.12}$$

This methodology can be implemented using Monte Carlo sampling of the model parameters' space.

Sobol's Method (Sobol, 1990a)

This is another variance-based method, which computes an ANOVA-like decomposition of the output variance with a Monte Carlo multidimensional integration of the main contributions of the parameters, the interactions and the higher order terms by the following equations:

$$f(\boldsymbol{X}) = f_0 + \sum_{i=1}^{n} f_i(X_i) + \sum_{1 \le i \le j \le n}^{n} \left[f_{ij}(X_i,\, X_j) + \cdots + f_{1\,2\ldots n}(\boldsymbol{X}) \right] \tag{2.13}$$

and assuming that $f(\boldsymbol{X})$ is squared-integrable we obtain to the next expression:

$$Var(Y) = \int_0^1 f^2(\boldsymbol{X}) d\boldsymbol{X} - f_0^2 \tag{2.14}$$

Then we perform similar tasks as in FAST using the first order sensitivities (S_i) to quantify the contribution of the individual parameters on the output variables, and the total sensitivities for the input parameters of the model (S_{ti}) to account for the effect of the correlations on the output variables. The measurements are defined as "the fraction of related partial variances to the overall variances" (Rahul, 2012).

The main disadvantage presented with this method is the computational challenge to estimate the integral in equation 2.14. However, the method is advantageous to the FAST method because of its ability to account for higher-order terms in the development of the variance series (Saltelli and Sobol, 1997).

2.3.3 Graphical Global Sensitivity Method for Cellular Network Dynamics

Biochemical systems are represented by highly regulated networks of large number of biochemical reactions. Correspondingly, the mathematical models needed to describe these systems are complex, highly nonlinear and involve a large number of parameters. To make these models robust it is imperative to analyze their behaviour with respect to parametric uncertainty and it is important to assess which parameters have the most effect on the model outputs. As discussed earlier, local sensitivity analysis cannot generally describe the complexity and correlated nature of these models. Correlation among model parameters is especially pervasive in biological systems due to the particular form of kinetic expressions, e.g. due to parameter correlation in Michaelis-Menten expressions, and to the highly interconnected nature of metabolic networks Therefore, it is necessary to analyze and understand how these correlations between the large numbers of parameters affect the sensitivity of outputs variables with respect to changes in model parameters. As explained in section 2.3.2 global sensitivity is needed in order to study the effect of simultaneous changes in model parameters on all the system's outputs.

Rand (2008) proposes a global parametric sensitivity analysis of a dynamic cellular network by means of two analytical tools: a *Sensitivity Heat Map* and a *Parameter Sensitivity Spectrum.* An added advantage of these tools is that their outcomes can be plotted thus permitting easier visualization of the sensitivity. Rand's method can be generalized for oscillatory and dynamical models.

The methodology proposed by Rand (2008) consists of a number of steps as follows:

i- A matrix is constructed with elements calculated from the partial derivatives of each output variable with respect to each input parameter at the sampling intervals of a given run. The key is that the partial derivatives of all outputs with respect to all parameters are considered together in order to account for correlations among parameters.

$$\boldsymbol{M} = \begin{bmatrix} \frac{\partial g_1}{\partial \eta_1}(t_1) & \frac{\partial g_1}{\partial \eta_2}(t_1) & \cdots & \frac{\partial g_1}{\partial \eta_j}(t_1) \\ \frac{\partial g_1}{\partial \eta_1}(t_2) & \frac{\partial g_1}{\partial \eta_2}(t_2) & \cdots & \frac{\partial g_1}{\partial \eta_j}(t_2) \\ \vdots & & & \\ \frac{\partial g_1}{\partial \eta_1}(t_T) & \vdots & \ddots & \vdots \\ \frac{\partial g_2}{\partial \eta_1}(t_1) & \vdots & & \vdots \\ \vdots & & & \\ \frac{\partial g_m}{\partial \eta_1}(t_T) & \frac{\partial g_m}{\partial \eta_2}(t_T) & \cdots & \frac{\partial g_m}{\partial \eta_j}(t_T) \end{bmatrix}$$

(2.15)

ii- The matrix is normalized to ensure independence from the choice of the time interval.

$$\boldsymbol{M}_1 = \sqrt{\frac{\Delta t}{T}} * \boldsymbol{M}$$

(2.16)

iii- "Thin Singular Value Decomposition" is applied to decompose the control coefficient matrix into a product of three matrices. Where $\boldsymbol{U}$ contains the orthogonal unit vectors, $\boldsymbol{\sigma}$ is a diagonal matrix with non-negative numbers, and $\boldsymbol{V}^T$ is an orthogonal matrix transposed.

$$\boldsymbol{M_1} = \boldsymbol{U}\,\boldsymbol{\sigma}\,\boldsymbol{V}^T$$

(2.17)

iv- The sensitivity heat map is obtained by multiplying the diagonal matrix ($\boldsymbol{\sigma}$) times the maximum values of the absolute value of the orthogonal matrix transposed ($\boldsymbol{V}^T$), and then times the matrix with the orthogonal unit vectors ($\boldsymbol{U}$).

$$f_{i,m}(t) = \sigma_i\,(\max_j |W_{ij}|)\,|U_{i,m}(t)|$$

(2.18)

v- The parametric sensitivity spectrum is achieved by applying the logarithm base ten to the absolute value of the product of the diagonal matrix ($\boldsymbol{\sigma}$) times the orthogonal matrix transposed ($\boldsymbol{V}^T$=W).

$$S_{ij} = \sigma_i \, V_{ij}^T = \sigma_i \, W_{ij}$$

(2.19)

$$\log_{10}\left|S_{ij}\right|$$

(2.20)

The sensitivity heat map shows us the contribution of the combination of parameters for each output variable over time. The method is analogous to the Principal Components Analysis technique where each principal component contains a relative contribution of an input to the overall variability of an output. In this method the principal components computed from the SVD decomposition compute the relative contribution of each parameter to each principal component to explain the change in the output. And the parametric sensitivity spectrum helps us identify the input parameters that are significantly large within each principal component involving a particular combination of parameters' contributions.

The results obtained with this approach provide a global presentation of the sensitivity analysis by means of a fundamental observation of the variables and the inputs. The sensitivity heat map and the parametric sensitivity spectrum can be plotted to facilitate visualization. Also, the computational time used for this approach is relatively small and easy to carry out as compared to other global sensitivity methods that require extensive Monte Carlo sampling.

The main limitations of the method is that it cannot take into account stochastic statistical distributions of the input parameters other than normal and most importantly that the analysis is applicable for a particular set of operating conditions. For example, if a batch process is analyzed Rand's method only considers the output values corresponding to this batch operation. However, in view that our objective was to perform the sensitivity analysis repetitively within an optimization search, Rand's method was chosen in this thesis due to its relative computational efficiency as compared to other global sensitivity methods that require Monte Carlo sampling.

2.4 PARAMETER ESTIMATION

Parameter estimation is a process system engineering activity whose goal is obtaining the parameter values of the model by matching the values calculated from the mathematical approximation function to the set of real data measurements. This approximation relies on assumptions related to statistical distribution of the parameters. Parameters may be time invariant or time-varying. Also, parameter values may be time invariant along a particular run of a batch process but may differ in value for different batch runs. The estimation computes numerical values for the parameters based on the data obtained from the observable variables (Dochain, 2002). The parameter estimation procedure generally provides a mean value of the parameter and an associated confidence interval or statistical distribution. In the case of time invariant parameters the identified statistical distribution of the parameters results from noise in the output data, and model structure errors. On the other hand, for time varying parameters, the identified statistical distribution of the parameters results from the combined effects of time variation, noise and model structure error.

The function that is used to calculate the values from the parameters by considering these as stochastic variables is named an estimator. The result from the estimator is called an estimate. The precision depends on the standard deviation of the estimate, because this is the measure of the errors caused by the calculation of the parameters and is affected by the fluctuations in the results. The bias is defined as “the deviation of the expectation of the estimate from the hypothetical true value of the parameter” (van den Bos, 2007). In this case the estimator is considered more and more accurate as the bias is reduced.

There are two main approaches for parameter estimation depending on the type of model used. Linear estimation is used to characterize approximately linear model functions, and nonlinear estimation, which is used to estimate the parameters from nonlinear model functions. The last approach is the most common in dynamic chemical and bio-chemical systems due to its inherent nonlinear behavior (Englezos and Kalogerakis, 2001; van den Bos, 2007).

2.4.1 Formulation

When choosing a parameter estimation procedure there are two main questions to be addressed: Which kind of model is more suitable for describing the process to be identified? And what is the objective function that is best suited for quantifying the fitting between model predictions and experiments?

Parameter estimation problems are generally formulated as optimization problems. In most cases the unknown parameters are obtained by solving an optimization problem, which involves minimizing or maximizing an objective function. This objective function is a measure of the gap between the data and the model (Bard, 1974; Seinfeld and Lapidus, 1974).

There are also two main assumptions that have to be considered during the formulation of the problem. First, it is assumed that the model structure is known and can potentially explain the data. And secondly, the solution of an optimization problem will result in a suitable parameter set where the parameter values do not significantly contradict physical sense or prior knowledge, e.g. a thermophysical property cannot be negative (Englezos and Kalogerakis, 2001).

The steps to follow for Parameter Estimation are (Englezos and Kalogerakis, 2001):

- Identify the structure of the model (linear or nonlinear)
- Determine the objective function to quantify the error between model predictions and data
- Choose the optimization method to minimize the objective function
- Determine the accuracy of the estimates
- Determine the adequacy of the model from the identified statistical distribution of the parameters

Validate the calibrated model with data that was not used for model calibration

The models to be used for parameter estimation are classified into two main groups: Algebraic models and Differential Equation Models.

Algebraic Model (Englezos and Kalogerakis, 2001) are given as follows:

$$y_i = f(x_i, k) + \varepsilon_i, \tag{2.21}$$

where y_i represents the dependent variables, x_i represents the independent variables, k the unknown parameters, and ε_i is the measurement errors.

Differential Equation Models (Englezos and Kalogerakis, 2001) are as follows:

$$\frac{dx(t)}{dt} = f(\boldsymbol{x}(t), \boldsymbol{u}, \boldsymbol{k}) \; ; \; \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{2.22}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{k}) + \boldsymbol{\varepsilon} \tag{2.23}$$

where $\boldsymbol{k}$ represents the parameter vector, $\boldsymbol{x}$ the vector with the state variables, $\boldsymbol{x_0}$ the vector with the initial conditions, $\boldsymbol{u}$ the vector with the manipulated variables, $\boldsymbol{y}$ the output variables vector, , $\boldsymbol{h}$ is a nonlinear function vector that relates the inputs to the output variables, and $\boldsymbol{\varepsilon}$ is the measurement error vector.

The objective function quantifies the distance between the model predictions and the data. The differences between output predictions and the data are often referred to as residuals and are mathematically given as follows (Englezos and Kalogerakis, 2001):

$$e_i = [y_i - f(x_i, k)]$$

(2.24)

where $f(x_i, k)$ represents the output evaluation by the model using the estimated parameter value.

According to Englezos and Kalogerakis (2011), parameter estimation methods can be classified into two main groups: explicit estimation and implicit estimation. In the explicit estimation class the output variables can be explicitly expressed as a function of the inputs and the model parameters. On the other hand in the implicit estimation the input, output and the parameters values are related to each other through an implicit function.

2.4.2 Parameter Estimation Methods

Several parameter estimation methods have been reported and only the most widely used ones are reviewed in the section. More information about parameter estimation can be found in the works of Bard (1974), Beck and Arnold (1977), Englezos and Kalogerakis (2001), and van den Bos (2007).

Least Squares

This is the simplest and most widely used method in parameter estimation. Its simplicity arises from the fact that it can be applied to any model without a priori knowledge about the probability distribution that characterizes its variables since this distribution is not used in the methodology. The three main assumptions used in this method are: i- the expected value for the error is 0; ii- homoscedasticity or same variance of the residuals; and iii- the covariance of the error is 0, meaning that the errors have no correlation between them (Brichoff et al., 1991). This method is especially useful for curve fitting problems. The method can be further classified into two classes: **Unweighted and Weighted Least Squares**.

The **Unweighted Least Squares** methodology does not take into account the different dimensions or the units of the measurements and the model. It involves a simple minimization of the sum of the squared residual values (Beck and Arnold, 1977).

$$\Phi(k) = \sum_{i=1}^{n} [y_i - f(x_i, k)]^2 = \sum_{i=1}^{n} {e_i}^2 \tag{2.25}$$

The **Weighted Least Squares** methodology compensates for large differences in magnitude between variables by multiplying the residuals by a weight factor (Beck and Arnold, 1977).

$$\Phi(k) = \sum_{i=1}^{n} b_i {e_i}^2 \tag{2.26}$$

Maximum Likelihood

The estimate obtained from the maximum likelihood is the value of the parameters that maximizes the likelihood function subject to equality and inequality constraints in the case where that value exists. This method is highly recommended when a large sample is available, because the variance of their estimates is the least compared to others. Maximum likelihood methodology also takes into consideration the distribution that the error follows, whereas the regular regression method does not. This way it maximizes the fitting between the data to an assumed statistical distribution of the errors. The price of its usage is the assumptions that have to be done (Bard, 1974; van den Bos, 2007; Pollock, 2003).

The main assumptions used in maximum likelihood estimation are (Pollock, 2003):

i- Expected value of the error is zero (*$E(\varepsilon_i)=0$*).

ii- Homoscedasticity (*$Var(\varepsilon_i)= \sigma^2$*)

iii- The errors are uncorrelated (*$Cov(\varepsilon_i, \varepsilon_j)=0$*)

iv- The error has a normal distribution

$$N(\varepsilon; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

(2.27)

v- The errors are independently distributed.

$$\prod_{t=1}^{T} N(\varepsilon; 0, \sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{t=1}^{T} \varepsilon^2\right)$$

(2.28)

The likelihood function of a sample is (Bard, 1974):

$$L(k, \psi) = p(y - f(x, k)|\psi) = p(\psi|k),$$

(2.29)

where $\boldsymbol{\psi}$ represents the distributions parameters.

$$\max_k L(k; \psi)$$

or

$$\max_k q(k; \psi)$$

$$q(k; \psi) = \ln L(k; \psi)$$

(2.30)

The logarithm of the likelihood function is often used as the optimization objective. The maximization of the likelihood function can sometimes help with parameter reduction problems and reduces the computational time (Bard, 1974) because the logarithmic operation transforms the product of terms into their summation. Since the logarithm is monotonic with respect to the argument the location of the maximum of the likelihood or the logarithm are identical. Due to the use of the logarithm, this method is also known as log-likelihood function (van den Bos, 2007).

Pseudomaximum Likelihood

This method uses the Maximum Likelihood equations for the specific case that the outputs errors are assumed to be normally distributed. This assumption serves to simplify the calculation of the parameters thus making this technique very popular (Bard, 1974).

Bayesian Estimation Methods

The estimation of a parameter depends on the probability density function used to characterize the parameters. This approach calculates parameters' estimates based on the minimization of a risk function of the parameters model (Lehmann and Casella, 1998; Vaseghi, 2000). In contrast to the methods previously reviewed, this method requires prior information about the probability density function of the parameters.

The benefits of using this method are that the estimates are physically meaningful since they satisfy at least the a priori assumed parameter statistical distribution. Also, the method calculates a posterior density function of the parameters and based on this posterior probability it is possible to neglect parameters that result in low probability. One of the major problems with this method is that it requires a priori knowledge and the computational time due to the need to sample the parameter space (Levy, 2012).

Monte Carlo Methods

There are very useful methods for nonlinear models where the parameters and the outputs have non-normal statistical distributions. Since the Monte Carlo method does not require specific a priori assumptions on statistical properties of parameters and outputs it is of very general applicability. Accordingly, it can also be used to analyze the properties and accuracy of other estimation methods. The Monte Carlo Method involves the following steps (Beck and Arnold, 1977):

1. Define all the elements in the analysis: model equations, probability distribution for the errors and, if applicable, the prior distribution of the parameters.
2. Sample the independent variables from their corresponding distributions and calculate the corresponding outputs variables using the model equations.
3. Calculate the probability distribution function.
4. Estimate the parameter values from the samples of the parameters that were used in the previous step.
5. Reproduce the experiments by repeating steps 3 and 4 with a new different set of parameter values for *N* times.
6. Obtain the mean value of the parameters estimated in all the *N* times.

This simulation method can be used to obtain the estimates for any linear or nonlinear model. (Lehmann and Casella, 1998; Brooks et al., 1995). The main disadvantage of this method is that it requires large amount of computation for more accurate results.

2.4.3 Parameter Validation

Fisher Information

The objective of any parameter estimation procedure is to search for the parameter set that results in the best fit between model predictions with the real measurements. Often, one is interested in quantifying the amount of information that an output has about a parameter. This is important for assessing the confidence interval of the parameter identified from specific measured outputs. The Fisher information helps to solve this problem by measuring the quantity of information that each unknown parameter contributes to every random output variable that is characterized with a specific probability function. The Fisher information matrix represents the variance of the expected observable value information, which is referred to as the Fisher score (Lehmann and Casella, 1998). Additionally, the FIM can be used to discriminate the parameter set that best fits the model.

The Fisher score vector for a set of parameters is defined as follows (van den Bos, 2007):

$$\boldsymbol{s}_{\theta} = \frac{\partial \log\,[L(\boldsymbol{y};\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \tag{2.31}$$

where $\mathbf{y}$ represents the vector of the output variables of the model, $\boldsymbol{\theta}$ is the parameter vector and $L(\boldsymbol{y};\boldsymbol{\theta})$ is the likelihood density function of the output variables.

The Fisher information matrix is essentially a weighted covariance matrix (dispersion matrix), and can be defined as following:

$$\boldsymbol{F} = \boldsymbol{S}^{T}\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{S} \tag{2.32}$$

where $\boldsymbol{S}$ is the matrix of the partial differential equations of the outputs at different times with respect to each of the parameters, and $\boldsymbol{\Sigma}$ is the covariance matrix of the measured noise. Both matrices are given as:

$$\boldsymbol{S}(t_i) = \begin{bmatrix} \frac{\partial y_1(t_i)}{\partial \theta_1} & \frac{\partial y_1(t_i)}{\partial \theta_2} & \cdots & \frac{\partial y_1(t_i)}{\partial \theta_n} \\ \frac{\partial y_2(t_i)}{\partial \theta_1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial y_m(t_i)}{\partial \theta_1} & \cdots & \cdots & \frac{\partial y_m(t_i)}{\partial \theta_m} \end{bmatrix}$$

(2.33)

$$\boldsymbol{\Sigma} = \begin{bmatrix} var_1 & & & \\ & var_2 & & \\ & & \ddots & \\ & & & var_m \end{bmatrix}$$

(2.34)

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{S}(t_i)$$

(2.35)

Where equation (2.35) represents a summation of the FIMs defined for each time interval necessary to analyze an entire experimental run. The inverse of the diagonal elements of the FIM matrix provides a lower bound for the covariance matrix of the errors in the parameter set (Walter and Prozanto, 1990) as follows defined as:

$$d_{ii} = F^{-\mathbf{1}}(i, i)$$

(2.36)

Confidence Intervals

The parameter estimate is an approximation of a true value. This estimate is very likely to differ from the actual true value due to the variability in the outputs and possible variability in parameters for time varying systems. When estimating a parameter value it is important to quantify by how much the estimated value may differ from the expected value. This can be done by estimating upper and lower bounds where the real value of the parameter is located between these bounds also referred to as confidence intervals. The confidence intervals are a measure of a probability of the occurrence of a particular value of a parameter. As the confidence interval gets smaller the confidence of the estimated mean of the parameter is larger (Beck, 1977; van den Bos, 2007). The confidence intervals for a parameter can be defined as a function of the inverse of the elements of the FIM (equation 2.36) as follows:

$$[\hat{\theta}_i \pm t_{n-p,\frac{\alpha}{2}}\sqrt{d_{ii}}] \tag{2.37}$$

where $t_{n-p,\frac{\alpha}{2}}$ represents the t-distribution with n - p degrees of freedom and a confidence interval of 100 (1- α)% (Smith, 2014; Gallant, 1975). The confidence intervals are often used to describe parametric uncertainty in robust control and robust optimization problems.

Chapter 3

METHODOLOGY

In this chapter we propose a methodology for model calibration of a Dynamic Metabolic Flux model using parameter sensitivity analysis. A simple case study is used to illustrate this methodology. An overall parametric sensitivity of the model is quantified by the sum of the parametric sensitivity coefficients associated with each output. The coefficients are obtained from the global parametric sensitivity analysis proposed by Rand (2008) as reviewed in the previous chapter. The calibration of the dynamic metabolic flux model is then based on an optimization problem that involves the maximization of the overall parametric sensitivity measure.

3.1 DYNAMIC METABOLIC FLUX ANALYSIS ALGORITH

The formulation of the dynamic model from a metabolic flux network is presented here:

$$\max_{v_i} \sum v_i(t)$$

subject to:

$$\boldsymbol{S} * \boldsymbol{v} = \boldsymbol{b}$$

$$\boldsymbol{v}(t) \geq 0$$

$$\frac{d\boldsymbol{g}(t)}{dt} = \boldsymbol{S} * \boldsymbol{v} * X$$

$$\boldsymbol{g}(t) \geq 0$$

(3.1)

where $\boldsymbol{S}$ represents the stoichiometric matrix $m\ x\ n,\ m$ being the metabolite number and n the fluxes number, $\boldsymbol{v}$ represents the metabolic flux vector in the system, $\boldsymbol{b}$ is the vector of consumption or production rates of metabolites per unit biomass, and $\boldsymbol{g}(t)$ is the vector with the output concentrations function of time.

The maximization is solved at each time interval and linearization of the mass balance for each metabolite is defined as follows:

$$\boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k + \boldsymbol{S} * \boldsymbol{v} * \boldsymbol{X}_k \tag{3.2}$$

It is generally assumed that bacteria attempt to maximize its growth at all times and therefore the objective function in (3.1) is assumed to be the growth rate. The maximization of the objective in (3.1) representing the growth rate is subject to the mass balance constraints and the positivity of the fluxes since the correct direction of the reactions is assumed a priori based on information from available databases such as the KEGG (The Kyoto Encyclopaedia of Genes and Genomes). The optimization is solved at each specific time interval and the optimal values can be used to describe the metabolites consumptions and/or productions over time.

3.2 SENSITIVITY ANALYSIS FORMULATION

The Sensitivity Analysis is conducted with the method of Rand presented in the previous chapter as per the following steps:

Step 1 Construction of *M* matrix

The construction of the $\boldsymbol{M}$ matrix containing the sensitivity coefficients from the system is obtained from the partial differential derivatives of each metabolite with respect to each of the parameters of interest. A normalization of the parameters is done to avoid problems with magnitude orders.

$$
\boldsymbol{M} = \begin{bmatrix}
\frac{\partial g_1}{\partial \eta_1}(t_1) & \frac{\partial g_1}{\partial \eta_2}(t_1) & \cdots & \frac{\partial g_1}{\partial \eta_j}(t_1) \\
\frac{\partial g_1}{\partial \eta_1}(t_2) & \frac{\partial g_1}{\partial \eta_2}(t_2) & \cdots & \frac{\partial g_1}{\partial \eta_j}(t_2) \\
\vdots & & & \\
\frac{\partial g_1}{\partial \eta_1}(t_T) & \vdots & \ddots & \vdots \\
\frac{\partial g_2}{\partial \eta_1}(t_1) & \vdots & & \vdots \\
\vdots & & & \\
\frac{\partial g_m}{\partial \eta_1}(t_T) & \frac{\partial g_m}{\partial \eta_2}(t_T) & \cdots & \frac{\partial g_m}{\partial \eta_j}(t_T)
\end{bmatrix}
$$

(3.3)

$$
\eta_j = \frac{\partial k_j}{k_j} = \log k_j
$$

(3.4)

where *k* is the parameter to be examined, *η* is the normalized term of the parameter, and *g* is the model output at a specific time interval. The rows correspond to the output variables from the model at each time interval and the columns correspond to the parameters of the model.

Step 2 Matrix Normalization

$$
\boldsymbol{M}_1 = \sqrt{\left(\frac{\Delta t}{T}\right)} * \boldsymbol{M};\ \ 0 \leq \Delta t \leq T
$$

(3.5)

Step 3 Singular Value Decomposition Analysis

$$
\boldsymbol{M}_1 = \boldsymbol{U}\,\boldsymbol{\sigma}\,\boldsymbol{V}^T,
$$

(3.6)

$$\boldsymbol{V}^T = \boldsymbol{W}$$

(3.7)

$$\sum_{\boldsymbol{m=1}}^{\boldsymbol{n}} \int_{\boldsymbol{0}}^{\boldsymbol{T}} U_{i,m}(t)\, U_{j,m}(t)\, dt = \delta_{ij}$$

(3.8)

where U_i has the same size as $\boldsymbol{M}_1$ *(m x n)* and these two matrices are orthogonal to each other as defined in equation (3.8)**,** $\boldsymbol{\sigma}$ is a diagonal matrix with non-negative numbers *(n x n*) and are defined as the sensitivity singular values. $\boldsymbol{V}^T$ is an orthogonal matrix *(n x n)*. The decomposition is needed to obtain the Sensitivity Principal components U_{im} (t), key elements of this methodology (Rand, 2008).

Also a new set of transformed parameters λ is defined where each λ is related to the normalized model parameter η according to an orthogonal linear transformation showed in the next equation:

$$\lambda_{\boldsymbol{i}} = \sum_{\boldsymbol{j}} W_{\boldsymbol{ij}} \partial \eta_{\boldsymbol{j}},$$

(3.9)

The index i corresponds to the transformed parameters λ and the index j corresponds to the parameters.

Based on the equations above the deviations in the output variables can be expressed as a function of the singular values as follows:

$$\delta g(t) = \sum_i \lambda_i \sigma_i U_{im}(t) + O(\| \delta\eta \|^2).$$

(3.10)

Combining equations (3.9) and equation (3.10), we can obtain an equation for the parametric sensitivity of each output with respect to each parameter as follows:

$$S_{ij} = \sigma_i \, W_{ij} \tag{3.11}$$

$$\frac{\partial g}{\partial \eta_j} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{S}} S_{ij} \, U_{im}. \tag{3.12}$$

$U_{i,m}$ is defined as the unit vector corresponding to the transformed parameters λ_i and the output m, with a length of $\frac{T}{\Delta t}$.

Step 4 Sensitivity Heat Formulation

From equations (3.11) and equation (3.12), we define the a new unit vector function of time ($f_{i,m}(t)$):

$$f_{i,m}(t) = \sigma_i \, (\max_j |W_{ij}|) \, |U_{i,m}(t)| \tag{3.13}$$

This unit vector is crucial for the correlation analysis, because it will help identify the contribution that the transformed parameters λ_i have to the outputs m. To identify the correlations that are significant to the model we defined a threshold of 5% of the maximum value of $f_{i,m}(t)$, this way we identify also the outputs with a significant Sensitivity Coefficient and to neglect the ones that are not significant according to this criterion. This equation can be plotted as a function of time for clearer visualization of the sensitivity results.

Step 5 Parameter Sensitivity Spectrum Formulation

In order to perform the Sensitivity Spectrum equation (3.11), $S_{ij} = \sigma_i \, W_{ij}$, is used to graphically visualize the sensitivity of the system's outputs with respect to each specific parameter. Following equation (3.12). It is possible to assess the independent effect that S_{ij} has on the overall sensitivity of an output with respect to a parameter, since the $U_{im}(t)'s$ are orthogonal matrices. This property makes the S_{ij} matrix a good indicator of the effect of each parameter j on the transformed parameters λ_i has and with the use of the Sensitivity Heat see the effect on the output m in the partial derivatives $\frac{\partial g}{\partial \eta_j}$.

The Sensitivity Spectrum consists of the plot of the 3D bar graph $\log_{10}|\mathrm{S_{ij}}|$ function of the transformed parameters λ_i and the parameters *j*. From this plot it is possible to identify the parameters with the highest influence in the system.

3.3 THE PROPOSED USE OF PARAMETRIC SENSITIVITY ANALYSIS IN THE CURRENT RESEARCH

The idea in this research is to use the parametric sensitivity measures proposed by Rand to accomplish a number of tasks related to the identification of dynamic metabolic flux models as follows:

1- To identify the parameters that does not significantly affect the outcomes of the model.

2- To identify whether the ranges of parameters of high parametric sensitivity will have a significant effect on profit or process constraints.

3- To identify the relevance of the parametric sensitivity measures on the parameter estimation problem.

To accomplish tasks 1 and 2 we propose to maximize a lumped measure of the individual parametric sensitivity coefficients defined in the previous section subject to set based constraints identified from data as follows:

$$\max_{\theta} \sum \sum \left| S_{ij} * U_{i,m} \right|$$

$$S.t.\;\; Set\ based\ constraints\ from$$

$$the\ metabolic\ flux\ model$$

$$\boldsymbol{\theta^U \geq \theta \geq \theta^L}$$

(3.14)

The key idea behind this equation is to find the maximum of the possible parametric sensitivity for all the data available for a particular experiment. Then, using the Sensitivity Spectrum it is proposed to neglect parameters that are below certain threshold of significance. The rationale is that if for the maximum overall sensitivity it is possible to neglect certain parameters then it is expected that these parameters could also be safely eliminated for parameters' regions of lower parametric sensitivity (Task 1 above). Furthermore, it is important to assess whether the regions of maximal parametric sensitivity will affect a worst profit or a constraint since this will have a major effect on robust optimization outcomes based on the model under consideration (Task 2 above).

Finally, we will assess the relevance of the maximization in 3.14 for the purpose of parameter estimation as compared to a minimization of least squares approach. As part of this comparison we will propose a modification of the least squares criterion that uses parametric sensitivity information.

3.3.1 Set Based Constraints

The key idea behind set based constraints is to represent the data by convex sets. To this purpose, it is assumed that because of measurement noise or unmeasured disturbances, the metabolites' concentrations are always between upper and lower limits at each time interval for which data is collected. For example, typical set constraints for glucose concentrations for a batch of E-coli culture with 10% bounds are depicted in Figure (3.1). Metabolites are typically measured by HPLC, which exhibit generally a large variability of 10% or more (Dewasme et al., 2010). Also, repeatability of cell cultures is not high due to variations in growth media, size and quality of inoculum and heterogeneity of the cell culture. All these factors contribute to the data to be variable within bounds. Since experimental runs are generally scarce due to experimental costs, it is difficult to assign a particular probability to trajectories within these sets. Instead, a uniform probability is assumed for each trajectory within the sets (Findeisen et al, 2003).

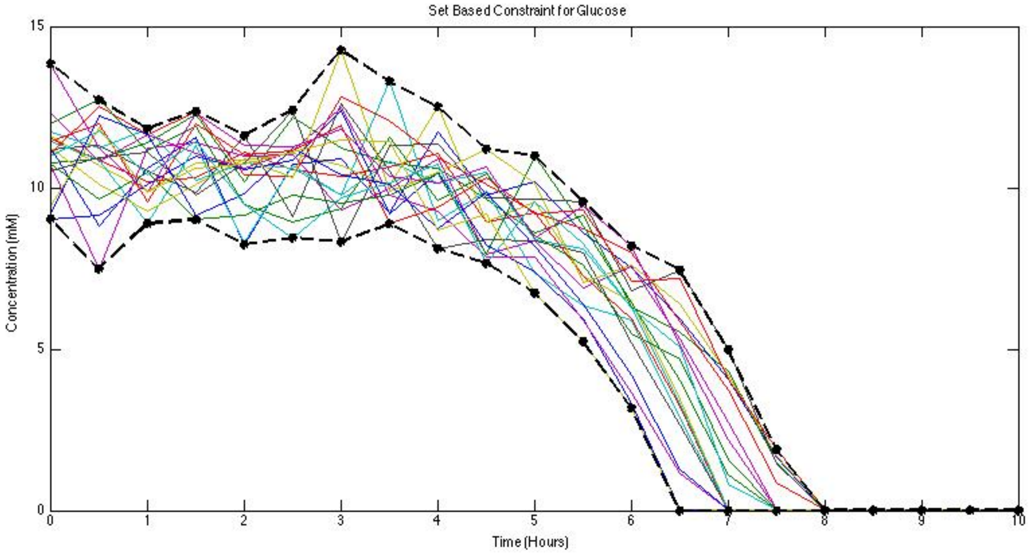

Figure 3.1. Set Based Constraint for Glucose

In the current thesis since the studies were conducted with simulations and experiments were not available, the Set Based Constraints where produced by simulations with the dynamic metabolic flux models. The combinations of two sources of variability were assumed for the data: unmeasured disturbances and measurement noise. To represent the unmeasured disturbance, several simulations were conducted by varying randomly the inputs parameter values between physically meaningful maxima and minima. To account for measurement noise, a random Gaussian noise was added to each of the output variables. The variations in the output variables obtained for simulations conducted with different parameters and noise were calculated and the maximum and minimum values at each time interval were assumed to define the bounds of the sets:

$$min\, \boldsymbol{Z}(t) \geq \boldsymbol{Z}(\widehat{\boldsymbol{\theta}}, t) \geq max\, \boldsymbol{Z}(t) \tag{3.15}$$

3.4 MODEL VALIDATION USING FISHER INFORMATION MATRIX AND CONFIDENCE INTERVALS

As reviewed in the previous chapter the elements of the Fisher Information Matrix can be used to calculate approximate confidence intervals on the parameters. Then, based on the magnitude of these intervals it is possible to assess the reliability of the parameter estimates. Smaller confidence intervals indicate higher confidence intervals. Also, if the confidence intervals are used to represent uncertainty for robust optimization purposes, then smaller intervals will generally result in less conservative optimization results. We will use this calculation to obtain the confidence intervals for the parameters estimated.

The formulation of the Fisher Information Matrix is defined as following:

$$\boldsymbol{F} = \boldsymbol{S}^T\, \boldsymbol{\Sigma}^{-1}\, \boldsymbol{S} \tag{3.16}$$

where $\boldsymbol{S}$ is the matrix of the partial differential equations of the outputs at one time respect to each of the parameters and it can be summed for all of the times analyzed, and $\boldsymbol{\Sigma}$ is the covariance matrix of the measured noise. These matrices are given as follows:

$$\boldsymbol{S}(t_i) = \begin{bmatrix} \frac{\partial y_1(t_i)}{\partial \theta_1} & \frac{\partial y_1(t_i)}{\partial \theta_2} & \cdots & \frac{\partial y_1(t_i)}{\partial \theta_n} \\ \frac{\partial y_2(t_i)}{\partial \theta_1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial y_m(t_i)}{\partial \theta_1} & \cdots & \cdots & \frac{\partial y_m(t_i)}{\partial \theta_m} \end{bmatrix} \tag{3.17}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} var_1 & & & \\ & var_2 & & \\ & & \ddots & \\ & & & var_m \end{bmatrix}$$

(3.18)

$$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{S}(t_i)$$

(3.19)

The FIM inverse matrix provides a lower bound for the covariance matrix of the errors in the parameter set (Walter and Prozanto, 1990). Where the element of the inverse FIM is defined as:

$$d_{ii} = \boldsymbol{F^{-1}}(i, i)$$

(3.20)

This equation is then used to obtain the two-sided confidence interval for the i-th estimated parameter, using the following equation:

$$[\hat{\theta}_i \pm t_{n-p,\frac{\alpha}{2}} \sqrt{d_{ii}}]$$

(3.21)

where $t_{n-p,\frac{\alpha}{2}}$ represents the t-distribution with n - p degrees of freedom and a confidence interval of 100 (1- α)% (Smith, 2014; Gallant, 1975).

Chapter 4

RESULTS AND DISCUSSION

4.1 CASE STUDY: Diauxic Growth of *E. coli*

The formulation of the dynamic model from a metabolic flux network for *E. coli* Diauxic Growth Model is presented here:

$$\max_{v_i} \ (v_1 + v_2 + v_3 + v_4)$$

subject to:

$$\boldsymbol{S} * \boldsymbol{v} = \boldsymbol{b}$$
$$v_i \geq 0$$

$$\frac{dg_{Ace}(t)}{dt} = S_{Ace} * v * X$$
$$\frac{dg_{Glu}(t)}{dt} = S_{Glu} * v * X$$
$$\frac{dg_{Oxy}(t)}{dt} = S_{Oxy} * v * X + k_L a * (0.21 - Oxygen)$$
$$\frac{dg_X(t)}{dt} = S_X * v * X$$
$$\boldsymbol{g}(t) \geq 0$$
$$\boldsymbol{g_0} = [10.8, 0.4 \ 0.21 \ 0.001]$$

$$\frac{dGlucose}{dt} \leq \frac{Vm * Glucose}{Km + Glucose}$$
$$\frac{dOxygen}{dt} \leq vo$$

(4.1)

where ***S*** represents the stoichiometric matrix $m\ x\ n$, m being the metabolite number and n the fluxes number, ***v*** represents the metabolic flux vector in the system, ***b*** is the vector of consumption or production rates of metabolites per unit biomass, km the substrate saturation constant, kLa the oxygen transfer coefficient, vo the maximal oxygen uptake, Vm maximal glucose uptake, $Acetate, Glucose, Oxygen, X$ are the concentration of the metabolites.

Nominal Parameter Values	Km [mM]	kLa [hr^{-1}]	vo [mmol/gdw hr]	Vm [mmol/gdw hr]
	0.015	**7.5**	**15**	**10**

Table 4.1. Nominal Parameter Values for Diauxic Growth Model (Mahadevan et al., 2002)

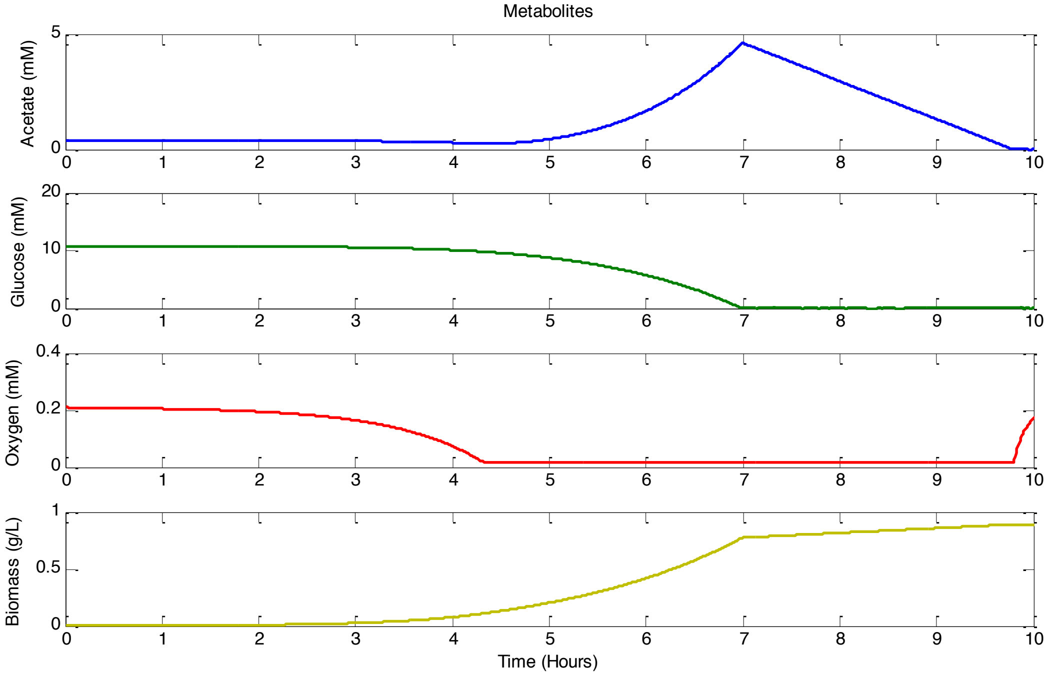

Figure 4.1. Metabolite Concentration over Time from Diaxuc Growth Model

4.2 SENSITIVITY ANALYSIS VALIDATION

The proposed methodology is based on the Jacobian matrix for the model outputs. The matrix is obtained by calculating the partial derivatives for each of the outputs with respect to each of the parameters. The parametric sensitivity analysis in this section was calculated at nominal values of the parameters specified by Mahadevan et al (2002). The outputs in the model were: Acetate, Glucose, Oxygen, and Biomass. The parameters used in the analysis were: substrate saturation constant (km), oxygen transfer coefficient (kLa), maximal oxygen uptake (vo), and maximal glucose uptake (Vm) and their nominal values are presented in the table 4.1.

The output of the model was obtained by running a simulation of a bacterial growth using the Dynamic Flux Balance Analysis (DFBA) of a diauxic growth for *Escherichia coli*, presented in the previous section. The evolutions of the metabolites over time during a batch culture are shown in Figure 4.1. In this case *E. coli* starts consuming the Glucose present in the media and produces Acetate as a secondary metabolite. Once the Glucose concentration is depleted, the uptake of the secondary carbon source, Acetate, starts taking place.

The partial derivatives matrix, ***M*** matrix, was calculated with respect to ten percent variations of the input parameters and was obtained for each output every two hours. The matrix thus formed consists of 20 rows corresponding to each of the four metabolites calculated every second hour for the total batch process that lasted ten hours, and 4 columns, which correspond to each of the parameters analyzed in the model.

In the next step a parametric sensitivity analysis of the model was conducted using the method of Rand (2008) reviewed in the Methodology chapter. First we normalized the elements of the ***M*** matrix by multiplying each one of them by the square root of the period of time where each output in measured, 2 hours, over the total time of the model's calculation, 10 hours, ($\sqrt{\Delta t/T}$). Then, the normalized ***M*** matrix was decomposed into a product of matrices by applying the thin singular value decomposition. The decomposition produces three matrices ***U***, with the same size as ***M*** *(20 x 4)* that contained the Sensitivity Principal components; $\boldsymbol{\sigma}$ with the size *4 x 4* that with the sensitivity singular values in decreasing order; and the square matrix V with the same dimension as the matrix containing the sensitivity singular values *(4 x 4)*. This last matrix helps to calculate the correlations and the contributions of the parameters due to the correlations among them as explained in the previous chapter.

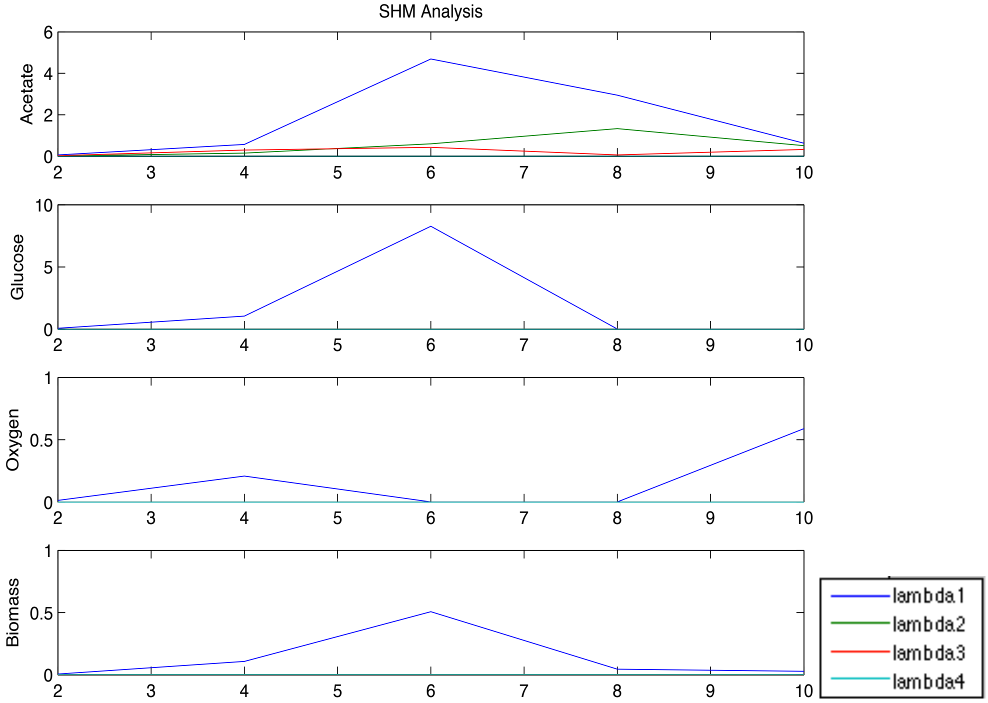

Figure 4.2. Analysis of the Correlation of the Parameters over time from Nominal Parameter set

Subsequently, the Control Coefficients $f_{i,m}$ were calculated from equation 3.13. This Control Coefficients relate the transformed parameters (λ_i), which characterize the parameter correlations, with the observable variables m. This way we capture the correlations among parameters with respect to the output variables as explained in the Methodology Chapter. In Figure 4.2 we plotted the Control Coefficients ($f_{i,m}$) over time for each of the output variables in the model (Acetate, Glucose, Oxygen, and Biomass). Here we can observe that the correlation variable (λ) with the highest influence is λ_1. This variable has influence on each of the output metabolites of the system. λ_2 is the second most important transformed parameter variable followed by λ_3, where these two latter variables have influence only on Acetate concentrations.

Then, a Parametric Sensitivity Spectrum (PSS) was used to identify which parameters from the correlations observed in the plots in Figure 4.2 have significant or negligible effect on the outputs. Based on the PSS it was possible to rank the parameters with the highest contribution to the model outputs, and also we found which parameters had no influence in the model. The calculation of the (PSS) was obtained from equation 3.11.

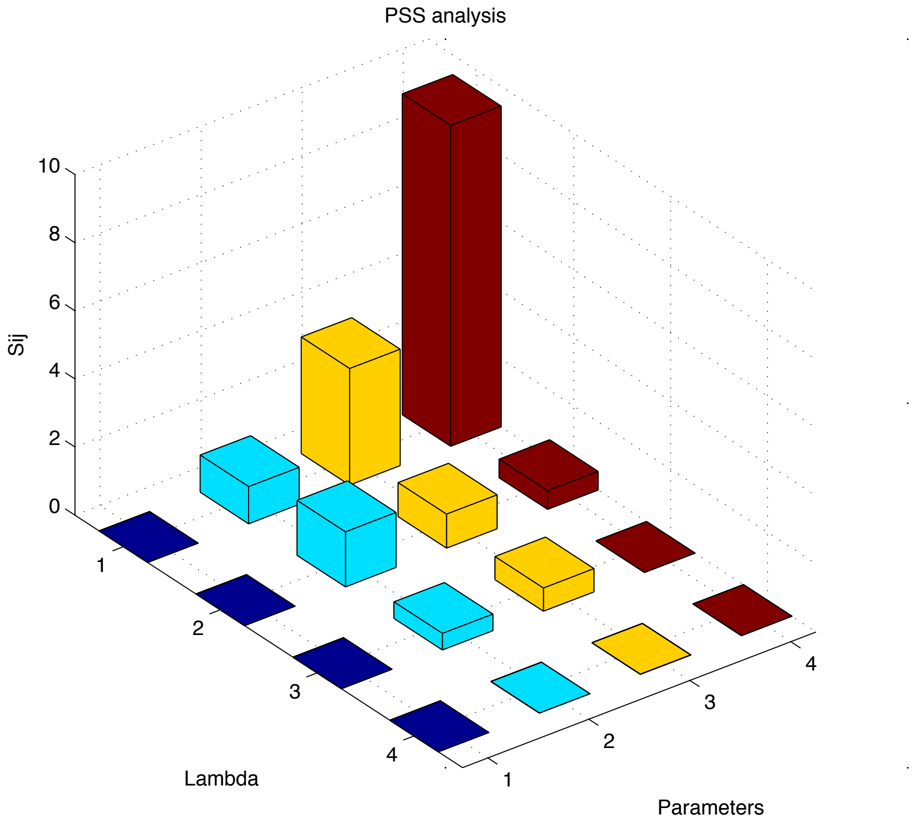

Figure 4.3. Sensitivity Spectrum Analysis on the 4 Nominal Parameters

Figure 4.3 shows the correlation variables (λ) on the y axis numbered from 1 to 4 and the parameters (*km, kLa, vo, Vm*) on the x axis from 1 to 4 respectively, both against their contribution value. This approach provides a link between the output variables and the parameters thus helping us to identify which parameters have more weight for each of the variables λ. The parameter 1 (*km*) is zero for each of the lambdas, meaning that the substrate saturation constant has no influence on the model outputs. Also we can observe that the parameter 4 (*Vm*) is the parameter with the highest contribution in lambda 1, meaning that the maximal glucose uptake constant is the parameter most influential in the model since the variable

λ_1 involves all the metabolites. The second most important parameter is the maximal oxygen uptake followed by the oxygen transfer coefficient.

The substrate saturation constant (km) determines the Glucose uptake rate and it is one of the elements in the denominator in the Michaelis-Menten equation 2.4. The small effect of this parameter can be explained by the fact that the value of this parameter reported in Mahadevan et al. (2002) is very small as compared to the glucose concentration during most of the batch and consequently it has negligible effect on the overall glucose uptake equation. In contrast, the maximal glucose uptake constant has the highest contribution to the model.

To verify the reliability of the parametric sensitivity analysis we introduced an additional artificial parameter (*kd*) with the goal of testing whether the analysis would lead to conclude that this parameter must be zero. The parameter was added into the numerator of the Michaelis-Menten equation (equation 4.2). The nominal value used for this artificial parameter to calculate its sensitivity was *0.1*.

$$v = \frac{Vm * [Glucose] + kd}{km + [Glucose]} \tag{4.2}$$

Sum of Sensitivity Coefficients (Objective Function)	**Case**
12.273	Nominal Parameter's Sum of Sensitivity Coefficients

Table 4.2. Sum of Sensitivity Coefficients of Nominal Parameter set

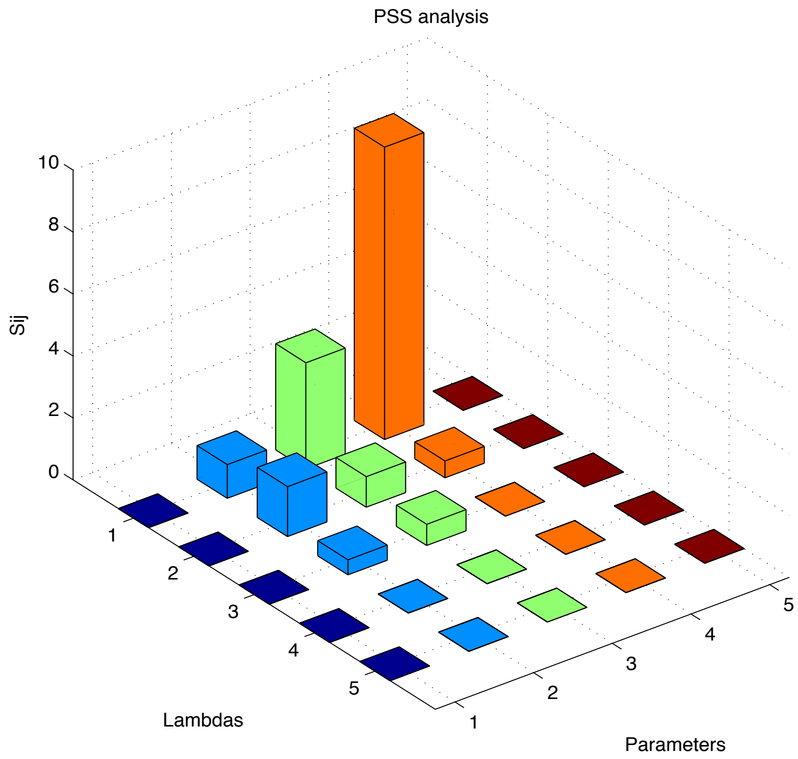

Figure 4.4. Sensitivity Spectrum Analysis on the Nominal Parameters and Artificial Parameter

The same sensitivity analysis was run in the similar conditions as the previous one and the analysis showed that the parameters with no influence in the system were substrate saturation constant, *km,* and the faked parameter, *kd,* which has no sensitivity due to its very small value compared to the other elements in the numerator. Once again, the parameter with the highest contribution is the maximal glucose uptake constant, *Vm*. In this way we were able to confirm that the sensitivity analysis in the model can correctly distinguish between parameters of significant versus insignificant effects on the outputs.

The goal in this last section was to identify the less significant parameters in the dynamic metabolic flux model and obtain the sensitivity coefficients of the relevant parameters in the neighbourhood of the parameter values reported by Mahadevan et al (2002) and the objective function to be maximized (Sum of the Sensitivity Coefficients) presented in table 4.2. In the next section the parameter values in the region of highest possible sensitivity are calculated from the optimization problem posed in 3.14 subject to set based constraints that describe data.

4.3 IMPLEMENTATION OF PARAMETER SENSITIVITY FOR PARAMETER ESTIMATION METHOD

The global sensitivity approach described above is used here to estimate parameters in the model in regions of high parametric sensitivity. Towards this goal we maximize the sum of the sensitivity contribution from the Parameter Sensitivity Spectrum. The maximization problem is given by equation 3.14.

The maximization was subject to set based constraints as explained in the methodology chapter. The upper and lower bounds of the sets were defined by positive or negative 10% changes with respect to the nominal values used in the study of Mahadevan et al. (2002). These 10% fluctuations in model parameters were used to represent unmeasured disturbances in the process. These disturbances may arise in a bioreactor due to changes in growth media, variability in the inoculums used to start each batch or errors in initial conditions (Dewasme et al., 2010).

In addition Gaussian noise was added to all the outputs to simulate sensor noise. Then, Set Based Constraints were obtained from simulations of the model given in equation 4.1 for different combinations of parameter values within the 10% bounds. From the resulting simulations we obtained a family of trajectories that are presented in the figures 4.5, 4.6, 4.7 4.8. These figures show the individual and combined contribution from parameters' changes and from noise. From these families of curves we obtain upper and lower bounds at different time intervals, which were used as constraints in the optimization problem.

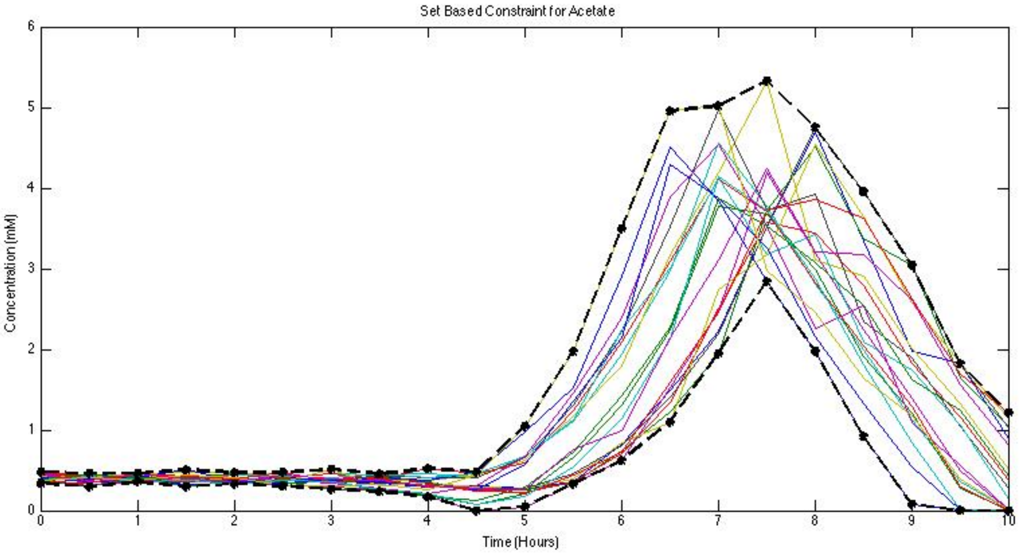

Figure 4.5. Set Based Constraints for Acetate

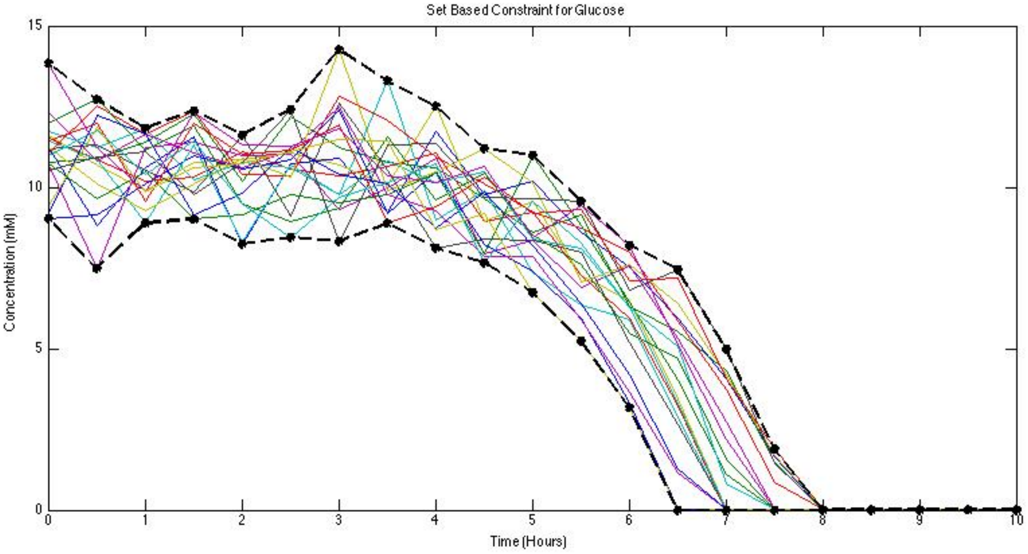

Figure 4.6. Set Based Constraints for Glucose

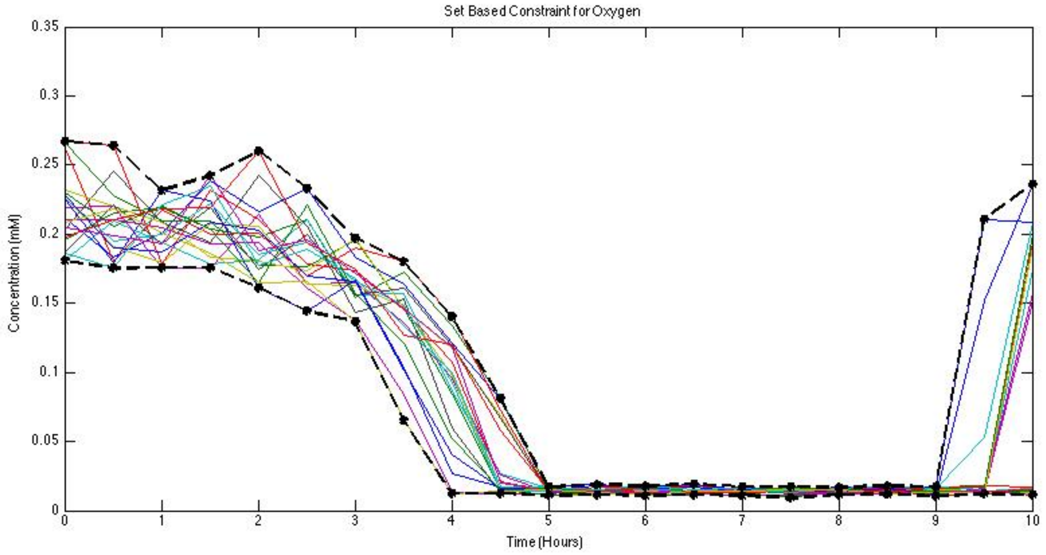

Figure 4.7. Set Based Constraints for Oxygen

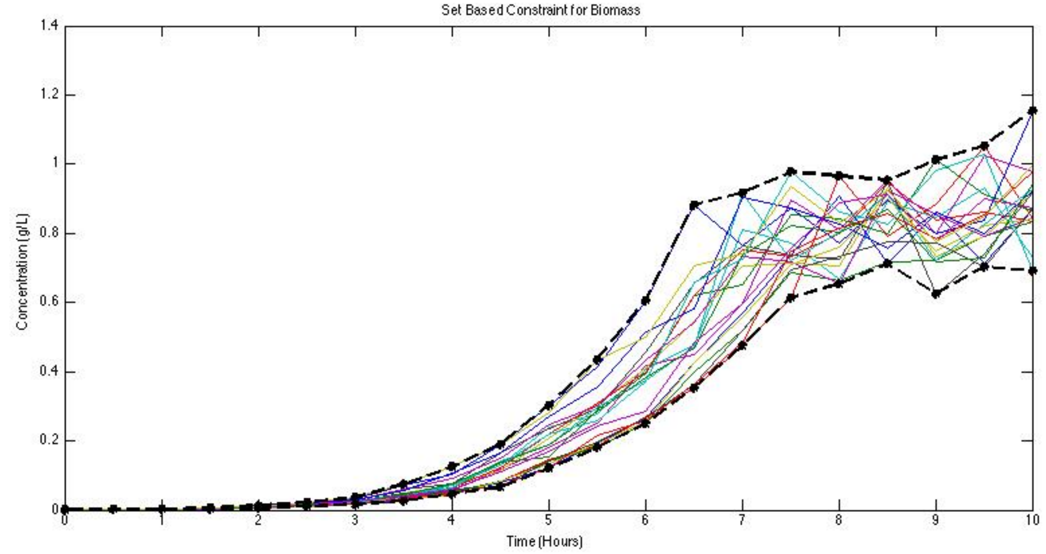

Figure 4.8. Set Based Constraints for Biomass

A first attempt was made to solve the optimization problem with the function Fmincon in Matlab. However, computational time was found to be a key challenge for solving the maximization problem in equation 3.14. The optimization requires repeated execution of the dynamic metabolic flux model. Each batch simulation of the dynamic metabolic flux model takes between 40 to 60 seconds to complete. Furthermore, the calculation of the ***M*** matrix requires the calculation of the partial derivatives for each parameter with respect to a deviation in each parameter. This requires executing twice the model, to generate each column corresponding to a deviation in each parameter. Thus, constructing the ***M*** matrix takes more than 90 percent of the total computational time of the optimization problem.

In order to reduce the computational time we tested several approaches to accelerate the calculation of the ***M*** matrix. The first idea was to omit the parameters with no effect on the model outputs, and avoid the calculation of its corresponding partial derivatives. In this way we could reduce the computational time by avoiding calculations of elements with insignificant contribution to the overall sensitivity.

Omitting the partial derivatives with respect to parameters with insignificant effect on sensitivity helped reducing the computational time to almost two fifth of the original computation time.

To further reduce the computational time we investigated the shape of the cost function. We found that the parametric sensitivity cost remains almost constant across a relatively wide range of parameter values and that is very nonlinear containing several similar minima. To address these issues, we implemented a logical loop of 10 iterations during which we did not change a subset of parameters that were found locally to be insensitive. For example if it was found at the beginning of a 10 iterations cycle that only 3 parameters out of 5 were important, we conducted the 10 subsequent iterations by maximizing with respect to these 3 parameters while the other 2 parameters were kept constant. At the end of the 10 iterations we optimized again with respect to the total number of parameters, find the ones of most effect on the outputs and proceed for the next 10 iterations by maximizing with respect to the parameters that their effect was found to be significant.

The results obtained using Fmincon always exhibit variability depending on the initial guess. This was expected in view that the cost seems to be very flat with several minima. To further address the flatness of the optimization cost surface and the occurrence of local minima we tested a discrete optimization approach where the parameter space was discretized among 1000 different combinations of model parameters and the corresponding parametric sensitivity was calculated.

An initial grid based on parameters' values corresponding to variations of $\pm 5\%$ and $\pm 10\%$ with respect to the nominal parameters reported by Mahadevan et al. (2002) and found in the table 4.1 was tested. Subsequently, the grid was further refined around the region of highest sensitivity found with the initial coarse grid. The refined grid was created by varying the parameter values by $\pm 2.5\%$ and $\pm 7.5\%$ around the best solution obtained from the coarse grid. The maximum sensitivity value and the parameter set for each of the cases is presented in the table below.

Km	**kLa**	**vo**	**Vm**	**Kd**	**Sum of Sensitivity Coefficients**	**Case**
0.015	7.500	15.000	10	0.1	12.273	Nominal Values
0.015	7.687	13.500	9	0.1	14.525	$\pm 2.5, 5\ , 7.5, 10\ percent\ variation$
0.015	7.125	13.500	9	0.1	14.206	$\pm 5\ , 10\ percent\ variation$

Table 4.3. Comparison of values obtained from discrete analysis by combinations

By pursuing this discrete optimization approach we were able to obtain better outcomes than with the Fmincon Matlab optimization function. We can also notice that for the two most influential parameters, the maximal glucose and oxygen uptake constants (Vm and vo) remain the same for the coarse and the fine grids and only the oxygen transfer coefficient (kLa) varied slightly between the two grids.

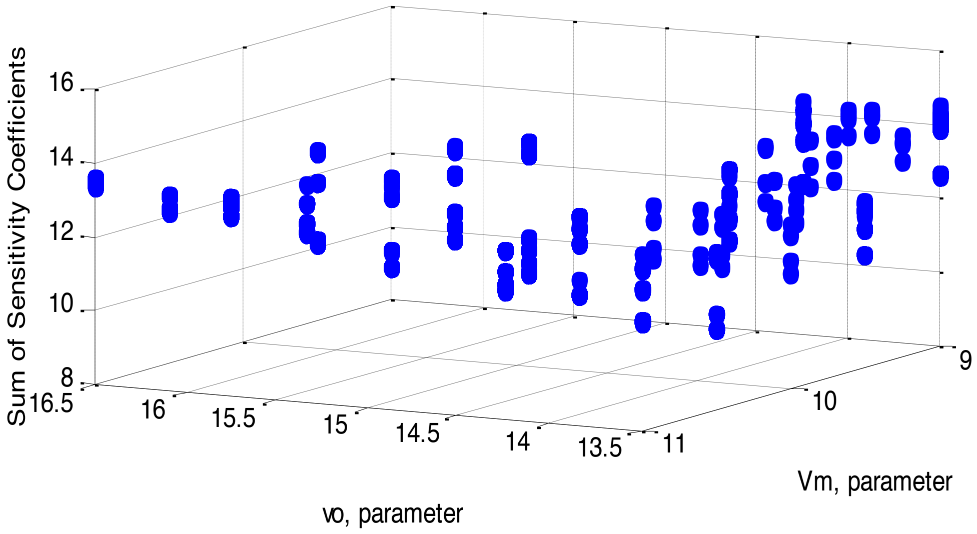

Figure 4.9. Graph Glucose and Oxygen uptake rates against Sum of Sensitivity

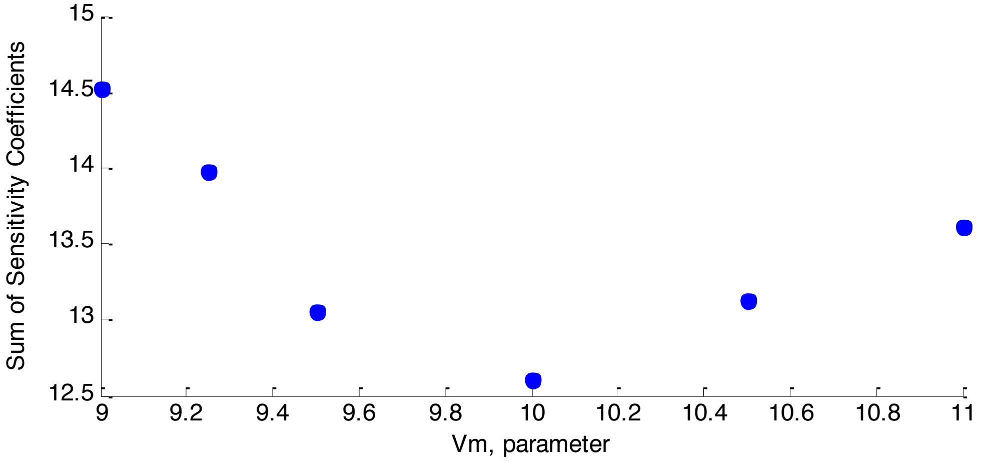

Figure 4.10. Graph of the Highest Sum of Sensitivity at each Glucose uptake rate variation

The plots in Figure 4.10 show how the model's sensitivity is mainly controlled by the changes of the glucose uptake rate (*Vm*). The sensitivity changes considerably by approximately 20% with respect to the sensitivity corresponding to the nominal value of Vm=10. This corroborates the sensitivity analysis results indicating that the model is mainly dominated by variability in this parameter.

4.4 MAXIMIZATION OF SENSITIVITY USING GENETIC ALGORITHM

Considering that the discrete approach was superior to a gradient based optimization fmincon) for maximizing the sensitivity, we consider using another maximization tool from MATLAB (The MathWorks Inc., Natwick, MA) based on a Genetic Algorithm that is also based on a discrete optimization approach. This algorithm calculates random initial guesses and selects combinations of parameters based on biologically motivated changes such as mutations and recombinations of parameter values to maximize a fitness function, which in this case is the parametric sensitivity. The outcomes obtained from the maximization of the sensitivity using the discrete optimization approach and the results obtained with Genetic Algorithm are compared in the table 4.4.

Km	**kLa**	**vo**	**Vm**	**Kd**	**Sum of Sensitivity Coefficients**	**Approach**
0.0150	7.6875	13.5000	9.0000	0.1000	14.5250	**Discrete Analysis**
0.0135	7.6091	13.5012	9.0000	0.1051	14.5916	**Genetic Algorithm**

Table 4.4. Comparison of Maximization the Sensitivity Approaches

As observed from table 4.4, the results for the maximization were slightly higher using Genetic Algorithm, but the results were generally similar. This proves that a discrete based optimization approach like Genetic Algorithm is more efficient for this problem as compared to continuous gradient based optimization approaches such as the one used by Fmincon.

Subsequently we compared the level of fitting in terms of the sum of square errors and the sensitivity coefficients obtained for two cases: i- parameters obtained from the maximization of the sensitivity as obtained from the solution of 4.1 and ii- parameters obtained from a least squares regression. The comparison between model predictions and the outputs corrupted by noise for these two sets of parameters are shown in Figure 4.11; and the results for the sum of square errors and the sensitivity coefficient for these two sets of parameters are presented in table 4.6 and the parameter sets corresponded to each method is presented in table 4.5.

Km	**kLa**	**vo**	**Vm**	**Kd**	**Curve Fitting Method**
0.0135	7.6091	13.5012	9.0000	0.1056	Sensitivity Analysis
0.0152	6.9099	15.7954	9.5332	0.0602	Least Squares

Table 4.5. Parameter sets for Sensitivity Analysis and Least Squares Fitting

Curve Fitting Method	**Sum of Squared Error**	**Sensitivity Coefficient**
Sensitivity Analysis	1084.0631	14.5914
Least Squares	659.5613	12.7766

Table 4.6. Sum of Squared Error and Sensitivity Coefficient for both Estimation methods

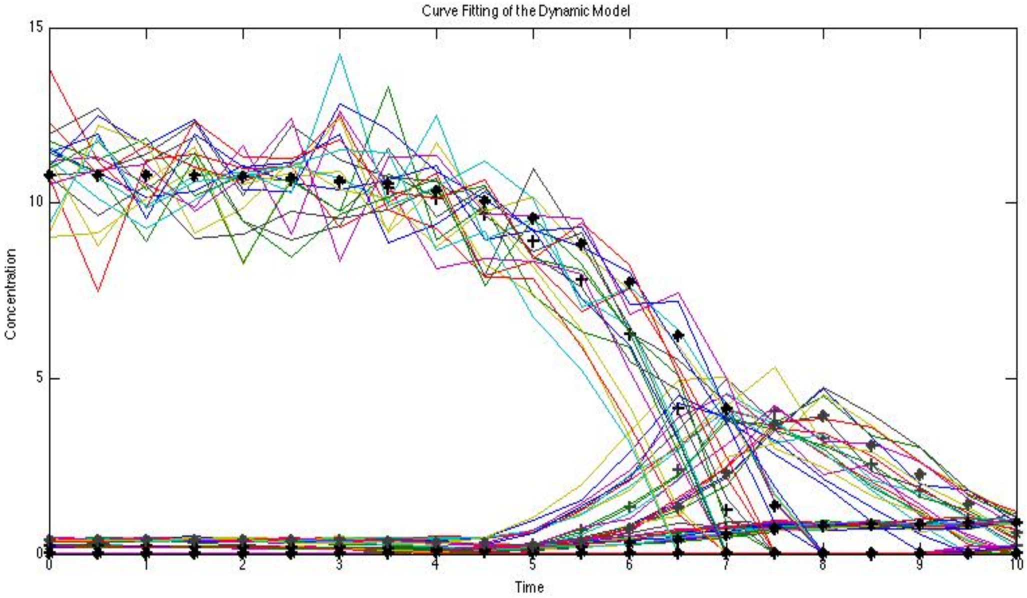

Figure 4.11. Curve fitting of Dynamic Model with Maximization of Sensitivity (‘*’) and Least Squares Fitting (‘+’)

It is evident from Figure 4.11 that the fitting using the least squares approach is better as compared to the fitting obtained with the parameters that maximize the sensitivity in terms of the sum of square errors. However the sensitivity coefficient for the Least Squares fitting is significant smaller thus resulting in larger confidence intervals as shown below.

The Fisher Information Matrix (FIM) analysis was implemented for the parameter sets of both estimation methods. The results from the FIM analysis of the most significant parameters in the two scenarios are presented in the table 4.7. It is clear that there is a substantial reduction in the magnitude of the confidence intervals: 2 percent for the oxygen transfer coefficient (*kLa*), 59 percent for the maximal oxygen uptake (*vo*), and 52 percent for the maximal glucose uptake (*Vm*).

	Sensitivity Analysis			Least Squares		
Parameters	**Mean**	**Variance**	**Confidence Intervals with $t_{\alpha,v}$=1.746**	**Mean**	**Variance**	**Confidence Intervals with $t_{\alpha,v}$=1.746**
kLa	7.6091	0.0016	±0.0701	6.8988	0.0017	±0.0711
Vo	13.5012	0.0019	±0.0761	15.8609	0.0053	±0.1276
Vm	9.0000	0.0007	±0.0447	9.5307	0.0024	±0.0854

Table 4.7. Mean, Variance, and Confidence Intervals of the significant parameters for both Parameter Estimation methods

One of the key applications of the confidence intervals in process systems engineering is to quantify uncertainty to be used for robust optimization. Robust optimization requires calculation of two key elements: i- a robust cost, i.e. a cost in the presence of uncertainty and ii- a robust gradient of the cost with respect to the decision variables in the presence of uncertainty.

Smaller confidence intervals are preferable since they will typically result in less uncertainty and less conservative optimization results. In this case study we assumed that the productivity is proportional to the amount of biomass. Although in the current study there was no a particular bio-product *E. coli* cultures will be typically used to produce a biomolecule of therapeutic interest. Generally, the amount of product will be proportional to the amount of biomass produced. Also, we assumed that the initial glucose concentration could serve as a possible decision variable to maximize the end of batch biomass. Thus, the gradient of interest for an optimization procedure is the gradient of biomass with respect to changes in initial glucose concentration. Then, to test the effect of uncertainty on a possible robust optimization formulation we calculated the effect of the parametric uncertainty as described by their confidence intervals on the biomass (cost) and on the gradient of biomass with respect to changes in initial glucose (gradient of cost with respect to decision variable).

First, the confidence intervals obtained from the FIM analysis s were used to generate a probability density function of the Biomass concentration based on the assumption that the parameters are normally distributed and the parameter variance is equal to the calculated confidence interval.

Then, to test the effect of the confidence intervals on the gradient of cost with respect to the initial glucose the Biomass concentrations at the time 8 hours for two inlet glucose concentration, 10.8 [mM] and 12.8 [mM] were calculated. Then the gradient of the cost with respect to the initial glucose was computed from the differences between the two final concentrations for Biomass over the two corresponding initial concentrations of glucose. A probability density function of the gradient was obtained by calculating this gradient for samples of normal distributions of parameters' combinations. The parameters were assumed to be normally distributed with a variance equal to the confidence intervals obtained from the FIM and with mean obtained from the least squares regression or from the maximization of the parametric sensitivity.

	Sensitivity Analysis		**Least Squares**	
Case	**Mean Biomass concentration at time 8 hours**	**Standard Deviation**	**Mean Biomass concentration at time 8 hours**	**Standard Deviation**
Gradient	0.0541	7.3912e-06	0.0538	7.7579e-05
Glu_0 = 10.8 [mM]	0.8028	2.2150e-06	0.8042	1.1045e-04
Glu_0 = 12.8 [mM]	0.9110	1.4514e-05	0.9119	1.0306e-04

Table 4.8. Mean and Standard Deviation of Biomass concentration at time 8 hours for both Parameter Estimation methods

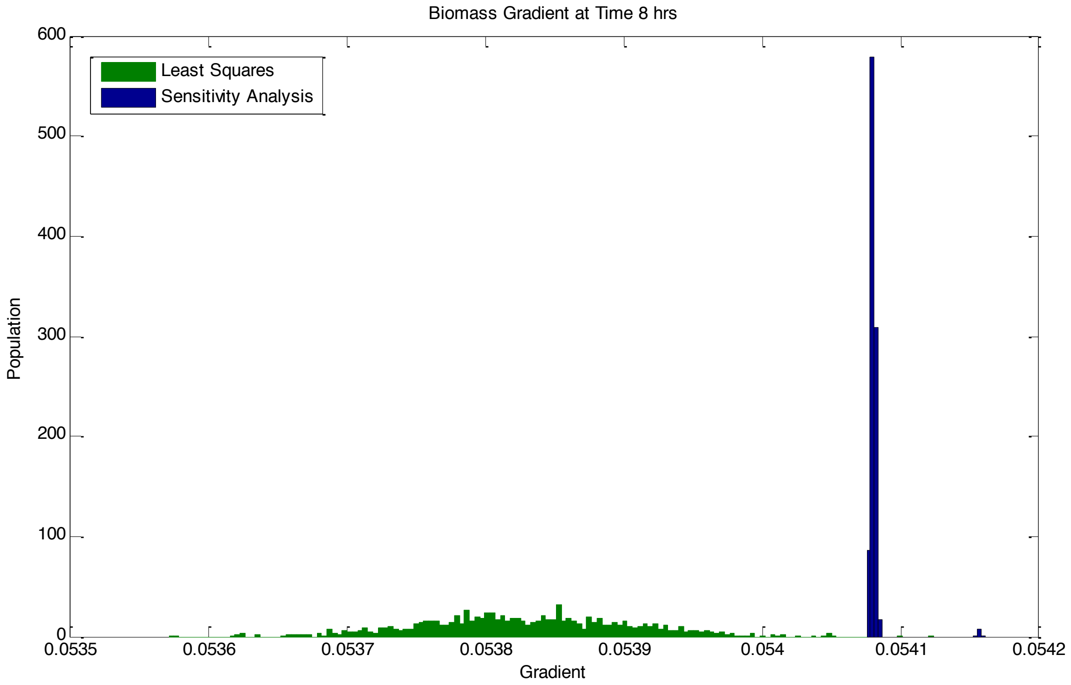

Figure 4.12. Normal Distribution Histogram of the Biomass Gradients at Time 8 hours for Least Squares and Sensitivity Analysis

As we can notice in the figures 4.12, 4.13, and 4.14, the normal distribution histograms for the biomass gradient based on the maximization of sensitivity show a significantly narrower curve compared to the ones obtained using sum of square errors. This difference between the two methods is also presented in the table 4.8 that shows that the standard deviations for all three cases using the maximization of sensitivity as proposed in this thesis are smaller. In both cases the standard deviation obtained with the maximal parametric sensitivity based solution is almost 40 percent smaller than the one based on sum of squared errors. This considerable reduction in the probability density function of the gradients of the cost shows the potential of working with parameters based on maximal sensitivity for reducing the conservatism of a robust optimization solution.

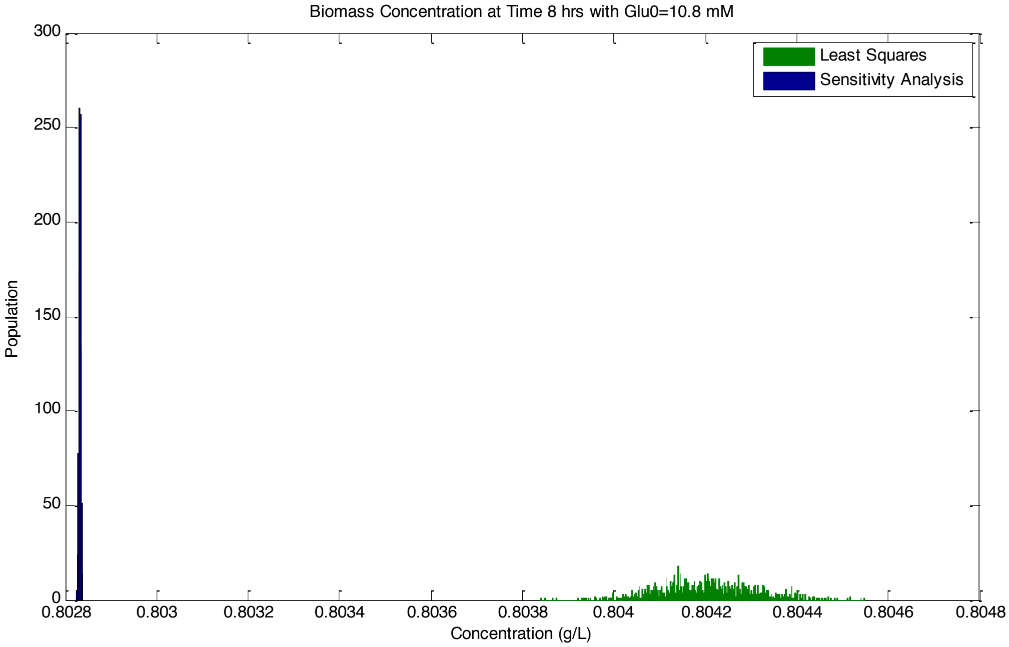

Figure 4.13. Normal Distribution Histogram of the Biomass Concentrations at time 8 hours with Initial Concentration of Glucose 10.8 (mM) for both Methods

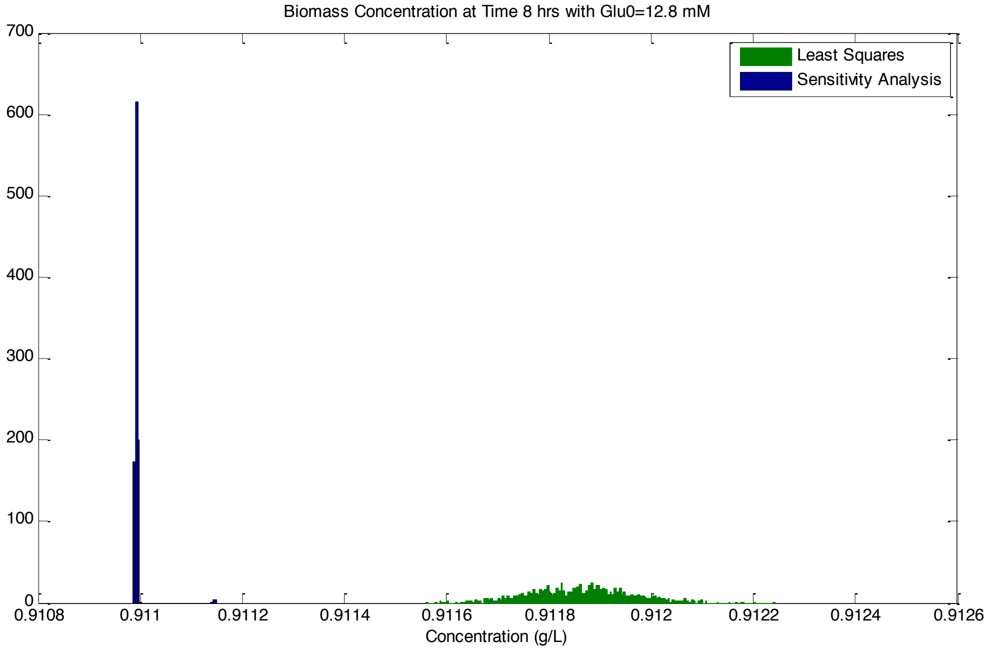

Figure 4.14. Normal Distribution Histogram of the Biomass Concentrations at time 8 hours with Initial Concentration of Glucose 12.8 (mM) for both Methods

4.5 ASSESSMENT OF SIGNIFICANCE OF MODEL PARAMETERS BASED ON MAXIMIZATION OF PARAMETRIC SENSITIVITY

Sum of Sensitivity Coefficients (Objective Function)	**Case**
12.2730	Nominal Parameter's Sum of Sensitivity Coefficients
14.5914	Maximized value of the Sum of Sensitivity Coefficients

Table 4.9. Sensitivity Coefficients of Nominal Parameter set and Maximized Parameter set

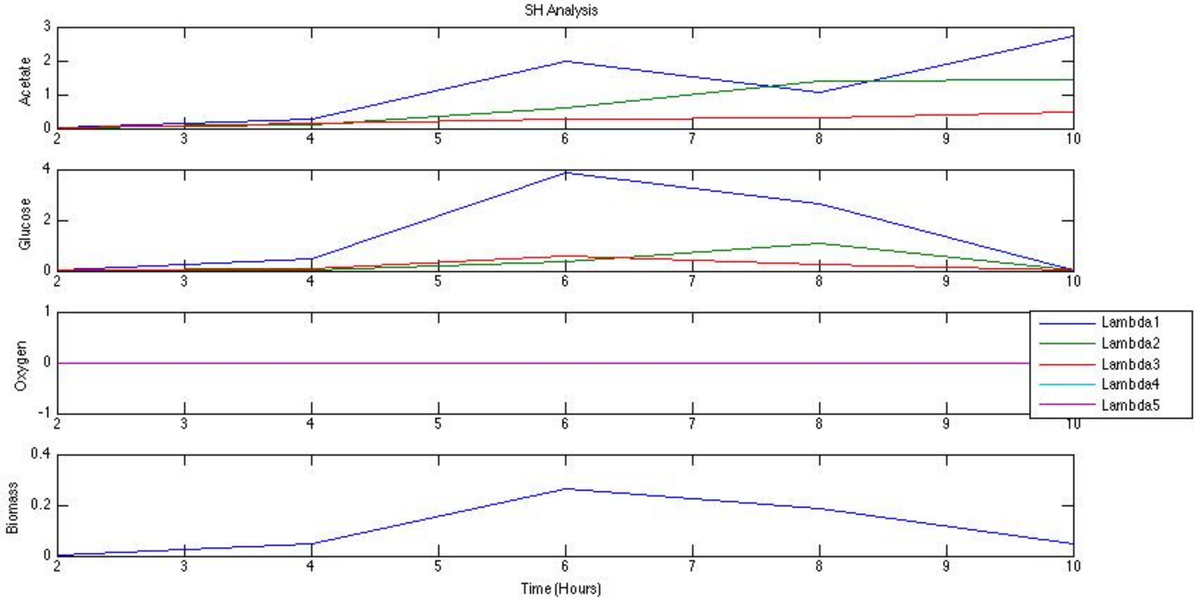

Figure 4.15. Analysis of the Correlation of the Parameters over time

In this section we consider again a set of 5 parameters (*km, kLa, vo, Vm, and kd*) where kd is an artificial parameter that is introduced to test the ability of the parametric sensitivity analysis to eliminate this redundant parameter. In this case the set based constraints were created with the actual model of Mahadevan without the parameter kd but the model used to maximize the sensitivity included kd. In Figure 4.15 we plotted the Control Coefficients ($f_{i,m}$) over time for each of the output variables in the model (Acetate, Glucose, Oxygen, and Biomass) using the parameter set obtained with the Maximization of the sensitivity method proposed in section 4.3. We present in table 4.9 a comparison of the Sum of Sensitivity Coefficients from the result of the sensitivity analysis without performing any maximization, as presented in section 4.2, and the objective function maximized using the methodology proposed in this thesis and which results are shown in section 4.3. Here it is possible to observe again that the correlation variable (λ) with the highest influence is λ_1 followed by λ_2 and λ_3 , where these latter two variables have influence only on Acetate concentrations and a smaller influence in the Glucose Concentrations.

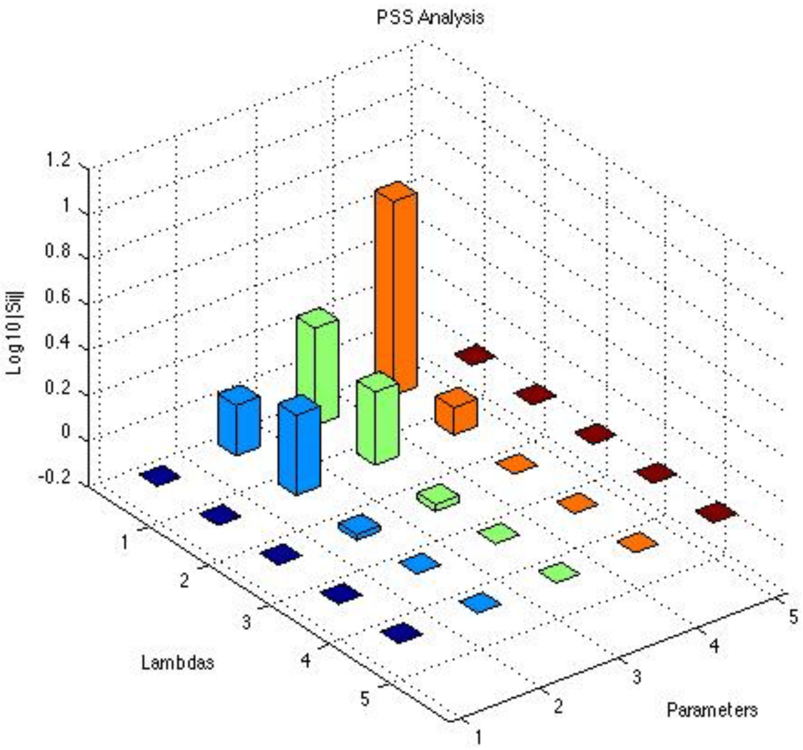

Figure 4.16. Sensitivity Spectrum Analysis on the 5 Maximized Parameter set (*km, kLa, vo, Vm, and kd*)

Figure 4.16 shows the correlation variables (λ) on the y axis numbered from 1 to 5 and the parameters (*km, kLa, vo, Vm, and kd*) on the x axis from 1 to 5 respectively, both against their contribution value. This plot can be used to rank the significance of the contributions of the parameters on the model. The parameters 1 and 5 (*km and kd*) have zero contributions for each of the lambdas, meaning that the substrate saturation constant and the artificial parameter included to validate the parametric sensitivity analysis have no influence on the model outputs. Also we can observe that the parameter 4 (*Vm*) is still the parameter with the highest contribution in lambda 1, meaning that the maximal glucose uptake constant is the parameter most influential in the model since the variable λ_1 is present in almost all the metabolites outputs. The second most important parameter is the maximal oxygen uptake followed by the oxygen transfer coefficient. We can observe also that for λ_2 the most influential parameters are parameter 2 and 3 (*kLa and vo*). However, these two parameters have only an effect in the Acetate concentration and a minor effect on glucose. This is expected due to the fact that in this model oxygen concentration controls the acetate uptake.

4.6 A HYBRID PARAMETER ESTIMATION APPROACH COMBINING PARAMETRIC SENSITIVITY AND THE SUM OF SQUARE ERRORS

The least squares methodology has been widely applied to estimate the parameter sets of many mathematical models used in Engineering. However this method is not fully accurate for nonlinear models since its accuracy is based on statistical assumptions that are only correct for linear systems Since least squares assumed equal importance for the errors, due to nonlinear dependence between outputs to parameters some errors may be important than others. In fact changes in parametric sensitivity directly reflect the changes in error magnitudes around different parameter values For example, in regions of high parametric sensitivity the changes in the outputs due to changes in parameters are expected to be larger thus resulting in a better signal to noise ratio and consequently smaller confidence intervals. On the other hand the parameters' values in regions of high parametric sensitivity may result in larger sum of square errors. Clearly, for models where the outputs are linear with respect to the parameters, the parametric sensitivity is uniform for all possible values of the parameters. For this reason, in view that for the model under study the outputs are nonlinear with respect to the parameters, we propose to normalize the sum of square errors by the sensitivity by dividing the sum by the parametric sensitivity measure used in the current study. In this way we expect to achieve a better trade-off between parametric sensitivity to sum of square errors. The following equation (4.3) describes the method to be implemented and the estimated parameter sets for this approach is presented below.

$$\min_{\theta}\left(\frac{Least\ Squares}{Parametric\ Sensitivity}\right)$$

or

$$\max_{\theta}\left(\frac{Parametric\ Sensitivity}{Least\ Squares}\right)$$

$$S.t.\ \ Set\ based\ constraints\ from$$

$$the\ metabolic\ flux\ \mathrm{model}$$

$$\boldsymbol{\theta^U \geq \theta \geq \theta^L} \tag{4.3}$$

The results of the optimization in (4.3) are shown in table 4.10. It is evident that some of the parameters are closer to the values calculated by least squares as compared to the parameters obtained with the maximization of sensitivity. Also we were able to increase the sensitivity of the model as shown in table 4.11, which will lead to values in the parameter set with higher probability than the regular Least Squares approximation. However, there was no considerable reduction in the confidence intervals for the parameters calculated with the new methodology, as presented in table 4.12. The possible reason will be that in the ratios defined in optimization 4.3 the Sum of Squares is penalized higher as compared to the parametric sensitivity. A better trade-off could be achieved by summing up the sum of squares and the parametric sensitivity with weights (table 4.13) but this is left for future studies.

Case	**Parameter Set**				
	km	**kla**	**vo**	**Vm**	**kd**
Nominal Parameters	0.015	7.5	15	10	0.1
Least Squares	0.01520	6.9099	15.7954	9.5332	0.0602
Max PS	0.01350	7.6091	13.5012	9.0000	0.1056
Min (LS/PS)	0.016407	6.8861	15.1625	9.6584	0.1051
Max (PS/LS)	0.016384	6.8476	15.6614	9.5476	0.1055

Table 4.10. Parameter Sets of the Estimation Methods

Case	Sum of Squared Errors	Sensitivity Coefficient
SSE	659.5613	12.7766
Max PS	1084.0630	14.5914
Min (LS/PS)	661.9028	13.3261
Max (PS/LS)	660.4011	13.3143

Table 4.11. Sum of Squared Error and Sensitivity Coefficient for Estimation Methods

	Max (PS/LS)			**Least Squares**		
Parameters	**Mean**	**Variance**	**Confidence Intervals**	**Mean**	**Variance**	**Confidence Intervals**
kLa	6.8476	0.0017	±0.0713	6.8988	0.0017	±0.0711
Vo	15.6614	0.0052	±0.1264	15.8609	0.0053	±0.1276
Vm	9.5476	0.0024	±0.0869	9.5307	0.0024	±0.0854

Table 4.12. Mean, Variance, and Confidence Intervals of the significant parameters for the Maximization of the PS over the LS Method and Least Squares Fitting.

	Min (PS/LS)			Least Squares		
Parameters	**Mean**	**Variance**	**Confidence Intervals**	**Mean**	**Variance**	**Confidence Intervals**
kLa	6.8861	0.0028	±0.0935	6.8988	0.0017	±0.0711
Vo	15.1625	0.0051	±0.1258	15.8609	0.0053	±0.1276
Vm	9.6584	0.0007	±0.0478	9.5307	0.0024	±0.0854

Table 4.13. Mean, Variance, and Confidence Intervals of the significant parameters for the Minimization of the LS over the PS Method and Least Squares fitting.

Discussion on the relevance of parametric sensitivity analysis for parameter estimation

A key question is on what is the advantage of finding the maximal sensitivity region as compared to the Least Squares solution with respect to the parameters' estimates,

The possible advantages of searching for the maximal sensitivity region are as follows:

1- Results in the parameter with the highest probability provided that it is assumed that each trajectory within the set based constraints have equal probability to occur.

2- It is particularly important if the region of high Parametric Sensitivity affects a worst case in terms of optimization, e.g. the region of parametric sensitivity affects the lower bound in productivity.

3- The Maximization of the Parametric Sensitivity solution results in the smallest confidence intervals for the parameters and thus may be advantageous for robust optimization with respect to decision variables since it is less sensitive to uncertainty (smaller uncertainty).

Advantage 1: Parameters with the highest probability if the probability of the output's errors is taken into account.

Lets assume a simple nonlinear function $y = C * f(\theta)$ with high parametric sensitivity at particular values of the parameter as shown in figure 4.17

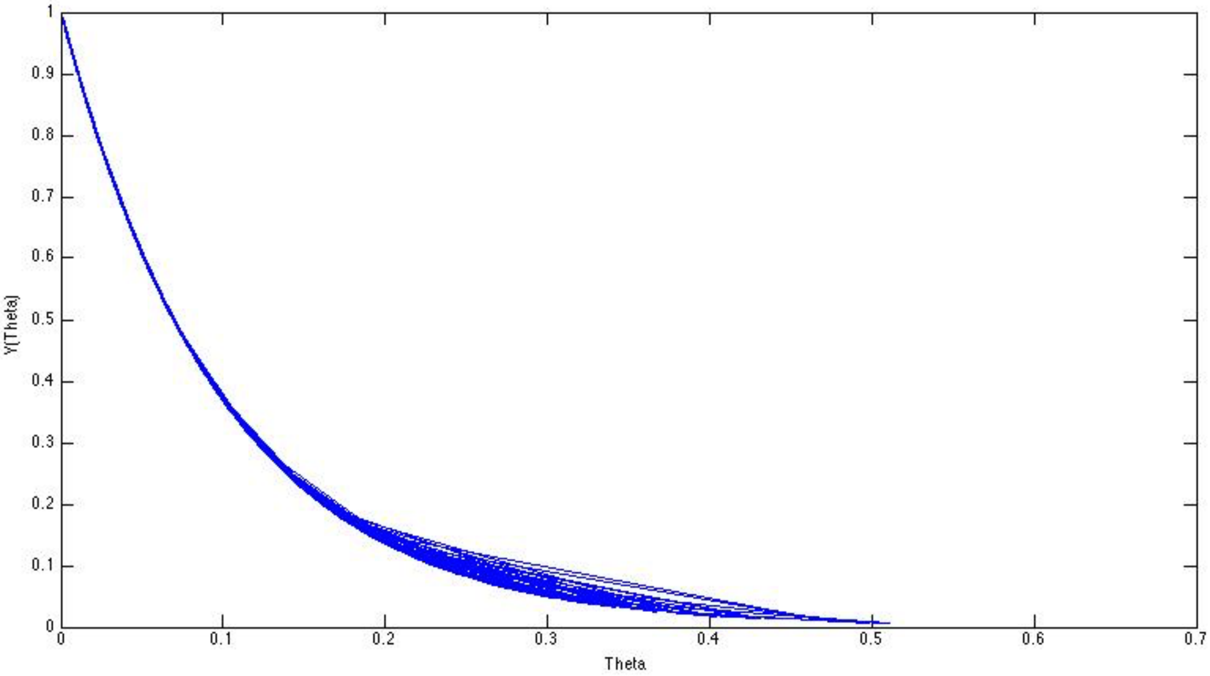

Figure 4.17 Exponential decay function of *Theta* (θ), which represents the sensitivity of the output approaching to high sensitivity parameter value (θ_0)

A key premise of least squares solution for resulting in a free bias solution is that the output errors are normally distributed. For a linear system if the parameters are normally distributed the output errors will be also normally distributed. This is not the case for nonlinear dependencies of the outputs with respect to the parameters that have varying Parametric Sensitivity around different parameter values, i.e. different values of the slope to the curve for different parameter values.

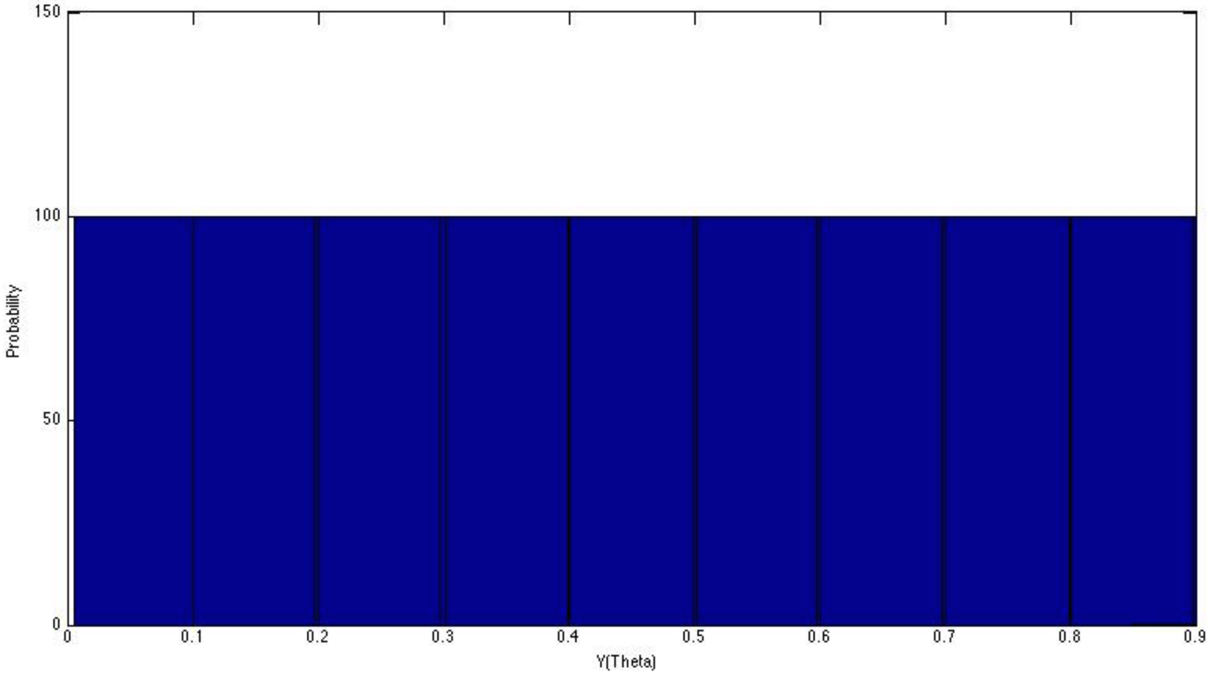

Figure 4.18 Uniform Probability of Occurrence of the Output ($y(\theta)$)

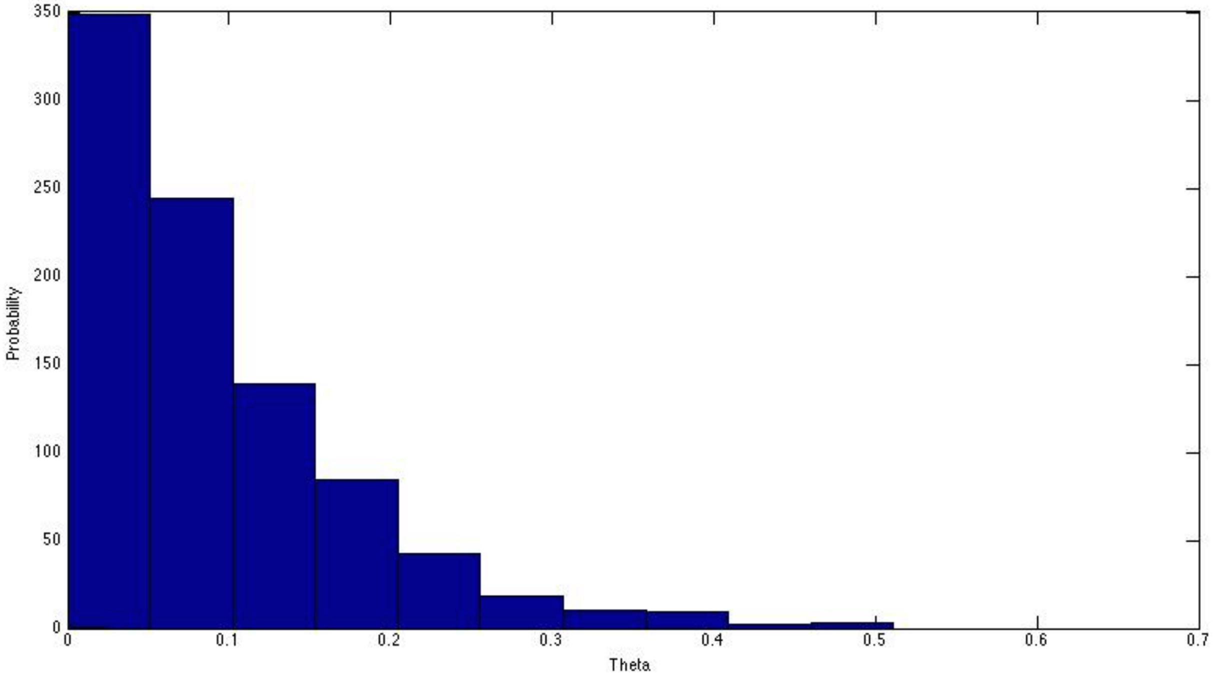

Figure 4.19 Parameter Probability Incidence

For example assume that the parameter is normally distributed about a θ_0 where θ_0 corresponds to a high Parametric Sensitivity region, i.e. distribution is centered on θ_0. The output will tend to spread in value within the high Parametric Sensitivity region and be more concentrated about a particular output value in low Parametric Sensitivity regions (figure 4.19). As a result of that the output distribution will not be normal and the Least Squares solution will find an average that will be away from the y corresponding to θ_0, i.e. the Least Squares solution will not result in the actual θ_0. The standard assumption when using set based constraints is that each trajectory within the sets has equal probability to occur (figure 4.18). In this case, the parameter estimates that correspond to the maximal parametric sensitivity will be the most probable parameters, i.e. smallest associated confidence intervals. For example, in Figure 4.17, if there is the same number of occurrences of s in the region of high sensitivity (e.g. $y(\theta)$ = 1) as in the region of low sensitivity (e.g. $y(\theta)$ = 0.1), the probability of occurrences of θ = 0.01 will be much higher than the probability of occurrences of θ = 0.4. The reason is that a much larger set of θ values can explain the changes around $y(\theta)$ = 0.1 than the range of parameter values explaining the changes around $y(\theta)$=1. Since generally output values will not have equal probability to occur within the sets, the maximal sensitivity will not be the most probable solution of the parameters. In that case we have proposed a hybrid approach where the Parametric Sensitivity values are normalized with the Sum of Square Errors.

Thus, the criterion for regression can be changed to a hybrid one resulting in $\min_\theta\left(LS\ (\theta)/PS(\theta)\right)$ or alternatively $\max_\theta\left(LS\ (\theta)/PS(\theta)\right)$ as presented in equation 4.3.

Advantage #2: region of high Parametric Sensitivity corresponds to output values that are relevant for optimization

One of the key applications for a model in chemical engineering is for performing model-based optimization. In the presence of parametric uncertainty it is often required to calculate a worst case. For example, for the *E. coli* case study one can maximize worst (lowest) productivity thus ensuring that the actual productivity will be always larger than the worst optimized case.

Identifying that the region of worst productivity corresponds to a region of high Parametric Sensitivity is very relevant since small changes of parameters, e.g. due to disturbances, will significantly affect the optimized bound.

This situation occurs in our case study where the region of maximal parametric sensitivity coincides with the region of lowest productivity. A robust optimization approach will attempt to maximize the worst (lowest) productivity.

Advantage 3: Robust optimization in the neighbourhood of an optimized worst bound.

Since the confidence intervals for a particular parameter are proportional to the inverse of the sensitivity (through the Fisher Information Matrix) then parameters' values within regions of high Parametric Sensitivity will have small confidence intervals.

Then, if one desires to do robust optimization by using the confidence intervals to quantify uncertainty, using parameters in the high PS region will have less associated uncertainty and will then result in less conservative predictions of the gradient of the cost with respect to changes in the decision variables used for optimization.

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS

This thesis investigated parametric sensitivity analysis as a tool for identifying a dynamic metabolic flux model. A novel approach to parameter estimation is proposed whereby the parametric sensitivity is maximized subject to set based constraints that are identified from data. Set based constraints are a very natural way to describe the data in biological systems where due to noise and disturbances, the experimental data is typically described by convex sets.

It is possible to conclude that the maximization of the parametric sensitivity of a model by means of a global sensitivity analysis subject to set based constraints served to find which parameters are not significant and also parameter values with small confidence intervals, i.e. with large level of confidence. The significance of the parameters was assessed in the regions of higher sensitivity as a way to assess significance in a worst case (highest sensitivity) situation. Although the current thesis dealt with a relatively simple example of *E. coli* growth, dynamic metabolic flux models may potentially involve a large number of metabolic reactions thus making the identification of these models to be a very challenging task. The parametric sensitivity method used in this thesis is particularly effective to deal with parameters that are correlated as occurring in typical dynamic metabolic flux models. Correlations are always present in metabolic flux models due to stoichiometric relations and to Monod kinetic structures where numerator and denominator parameters may exhibit correlations. Thus, considering each parameter independent from the other may lead to inaccurate results. For our case we were able to identify which correlations were affecting the concentration of each metabolite and since the analysis was applied at different time intervals it was possible to identify at which times they were more significant.

Also, we were able to identify from the correlations discussed above which parameters are more influential or which ones have no influence on the model. For example, the effect of substrate saturation constant (*km*) on the growth was found to be negligible within the range of values that were analyzed. This lack of significance is due to the low saturation constant value as compared to the glucose concentration levels occurring during the batch.

An additional advantage of the parameter estimation via a maximization of the sensitivity was to minimize the uncertainty of the parameters' estimates. Since higher parametric sensitivity translates into smaller confidence intervals, parameters' estimates in regions of high parametric sensitivity have larger probability (smaller confidence intervals). We have shown that estimating parameters in regions of high parametric sensitivity may be advantageous for robust optimization since the associated uncertainty for the estimated parameters is less conservative. This reduction of conservatism is especially important in terms of the gradients of the cost function with respect to changes in the decision variables that have to be used in robust optimization.

On the other hand the parameter estimates that maximize the sensitivity do not necessarily result in good fitting to data in terms of the sum of square errors. Thus, there is a trade-off between finding estimates of high probability via maximization of parametric sensitivity versus finding estimates with lower probability but better fitting via the minimization of the sum of square errors. To address this trade-off we have proposed a hybrid parameter estimation method where the parametric sensitivity normalized (divided by) the sum of square errors is maximized subject to the set based constraints. This hybrid approach is shown to achieve a trade-off between the two criteria, i.e. parametric sensitivity versus sum of square errors.

5.2 FUTURE WORK

A major challenge in the maximization of the sum of sensitivity coefficients was the computational time. The most time consuming step in this procedure is the calculation of the ***M*** matrix that is based on the partial derivatives of the outputs with respect to each parameter. A possible way to solve this difficulty would be to create a large look up table of partial derivatives and to apply a continuous gradient seeking optimization method based on interpolated values of derivatives from the look up table

In addition it is proposed to extend this methodology to a larger dynamic metabolic flux model. For example, a model for yeast growth is currently available that involves a 100 reactions. The methodology proposed in the current work will be instrumental to calibrate such model.

Finally it is proposed to use the current methodology to perform robust optimization based on a dynamic metabolic flux model. In the previous chapter it was shown that working with parameter estimates with smaller confidence intervals may be advantageous since it may reduce the conservatism of robust optimization solutions.

REFERENCES

Archer, G., Saltelli, A., & Sobol, I. (1997). Sensitivity measures, ANOVA-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation, 58*(2), 99-120.

Bard, Y., & Bard, Y. (1974). *Nonlinear Parameter Estimation* Academic Press Inc.

Beck, J. V., & Arnold, K. J. (1977). *Parameter estimation in engineering and science* James Beck.

Bischoff, W., Cremers, H., & Fieger, W. (1991). Normal distribution assumption and least squares estimation function in the model of polynomial regression. *Journal of Multivariate Analysis, 36*(1), 1-17. doi:http://dx.doi.org/10.1016/0047-259X(91)90087-I

Blower, S. M., & Dowlatabadi, H. (1994). Sensitivity and uncertainty analysis of complex models of disease transmission: An HIV model, as an example. *International Statistical Review/Revue Internationale De Statistique,* , 229-243.

Budman, H., Patel, N., Tamer, M., & Al‐Gherwi, W. (2013). A dynamic metabolic flux balance based model of fed‐batch fermentation of bordetella pertussis. *Biotechnology Progress, 29*(2), 520-531.

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software, 22*(10), 1509-1518.

Cukier, R., Fortuin, C., Shuler, K. E., Petschek, A., & Schaibly, J. (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I theory. *The Journal of Chemical Physics, 59*(8), 3873-3878.

Cukier, R., Levine, H., & Shuler, K. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics, 26*(1), 1-42.

Cukier, R., Schaibly, J., & Shuler, K. E. (1975). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III. analysis of the approximations. *The Journal of Chemical Physics, 63*(3), 1140-1149.

Dewasme, L., Richelle, A., Dehottay, P., Georges, P., Remy, M., Bogaerts, P., et al. (2010). Linear robust control of S. cerevisiae fed-batch cultures at different scales. *Biochemical Engineering Journal, 53*(1), 26-37.

Draper, N. R., & Smith, H. (1998). Fitting a straight line by least squares. *Applied Regression Analysis, Third Edition,* , 15-46.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* John Wiley & Sons.

Englezos, P., & Kalogerakis, N. (2000). *Applied parameter estimation for chemical engineers* CRC Press.

Findeisen, R., Imsland, L., Allgower, F., & Foss, B. A. (2003). State and output feedback nonlinear model predictive control: An overview. *European Journal of Control, 9*(2–3), 190-206.

Gallant, A. R. (1975). Seemingly unrelated nonlinear regressions. *Journal of Econometrics, 3*(1), 35-50.

Garrett, R. M., Rothenburger, S. J., & Prince, R. C. (2003). Biodegradation of fuel oil under laboratory and arctic marine conditions. *Spill Science & Technology Bulletin, 8*(3), 297-302.

GHIDERSA, B., Wörner, M., & Cacuci, D. (2003). Numerical simulation of bubble-train flow in a small channel of square cross-section. *Wissenschaftliche Berichte FZKA, 6759*, G. 1-G. 9.

Hamby, D. (1995). A comparison of sensitivity analysis techniques. *Health Physics, 68*(2), 195-204.

Helton, J. C., Iman, R., Johnson, J., & Leigh, C. (1986). Uncertainty and sensitivity analysis of a model for multicomponent aerosol dynamics. *Nuclear Technology, 73*(3), 320-342.

Helton, J. C., Iman, R. L., & Brown, J. B. (1985). Sensitivity analysis of the asymptotic behavior of a model for the environmental movement of radionuclides. *Ecological Modelling, 28*(4), 243-278.

Hjersted, J. L., Henson, M. A., & Mahadevan, R. (2007). Genome‐scale analysis of saccharomyces cerevisiae metabolism and ethanol production in fed‐batch culture. *Biotechnology and Bioengineering, 97*(5), 1190-1204.

Hjersted, J. L., & Henson, M. A. (2006). Optimization of fed-batch saccharomyces cerevisiae fermentation using dynamic flux balance models. *Biotechnology Progress, 22*(5), 1239-1248.

Iman, R. L., & Helton, J. C. (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Analysis, 8*(1), 71-90.

Iman, R. L., & Helton, J. C. and Campbell, J. E. (1981a). An Approach to Sensitivity Analysis of Computer Models: Part I - Introduction, Input Variable Selection and Preliminary Assessment. J. *Qual. Technol., 13,* 174-183.

Iman, R. L., & Helton, J. C. and Campbell, J. E. (1981b). An Approach to Sensitivity Analysis of Computer Models: Part II - Ranking of Input Variables, Response Surface Validation, Distribution Effect and Technique Synopsis. *J. Qual. Technol., 13,* 232-240.

Ingalls, B. (2013). Mathematical modelling in systems biology: An introduction.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research, 38*(Database issue), D355-60.

Kauffman, K. J., Prakash, P., & Edwards, J. S. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology, 14*(5), 491-496.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics, 5*(11), 826-837.

Kitano, H. (2002). Systems biology: A brief overview. *Science (New York, N.Y.), 295*(5560), 1662-1664.

Kitano, H. (2007). Towards a theory of biological robustness. *Molecular Systems Biology, 3*, 137.

Kitano, H., & Oda, K. (2006). Robustness trade-offs and host-microbial symbiosis in the immune system. *Molecular Systems Biology, 2*, 2006.0022.

Koda, M., Mcrae, G. J., & Seinfeld, J. H. (1979). Automatic sensitivity analysis of kinetic mechanisms. *International Journal of Chemical Kinetics, 11*(4), 427-444.

Lehmann, E. L., & Casella, G. (1998). Theory of point estimation (springer texts in statistics).

Liepmann, D., & Stephanopoulos, G. (1985). Development and global sensitivity analysis of a closed ecosystem model. *Ecological Modelling, 30*(1), 13-47.

Mahadevan, R., Edwards, J. S., & Doyle, F. J. (2002). Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical Journal, 83*(3), 1331-1340.

Michiels, W., Engelborghs, K., Roose, D., & Dochain, D. (2002). Sensitivity to infinitesimal delays in neutral equations. *SIAM Journal on Control and Optimization, 40*(4), 1134-1158.

Montgomery, D. C., & Runger, G. C. (2010). *Applied statistics and probability for engineers* John Wiley & Sons.

Niklas, J., Schneider, K., & Heinzle, E. (2010). Metabolic flux analysis in eukaryotes. *Current Opinion in Biotechnology, 21*(1), 63-69.

Nowruzi, K., Elkamel, A., Scharer, J. M., Cossar, D., & Moo-Young, M. (2008). Development of a minimal defined medium for recombinant human interleukin-3 production by streptomyces lividans 66. *Biotechnology and Bioengineering, 99*(1), 214-222.

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology, 28*(3), 245-248.

Pimm, S. L., Russell, G. J., Gittleman, J. L., & Brooks, T. M. (1995). The future of biodiversity. *Science, 269*(5222), 347.

Pollock, D. S. G. (2003). Recursive estimation in econometrics. *Computational Statistics & Data Analysis, 44*(1–2), 37-75.

Proudfoot, F. G., Hulan, H. W., & McRae, K. B. (1982). The effect of crumbled and pelleted feed on the incidence of sudden death syndrome among male chicken broilers. *Poultry Science, 61*(8), 1766-1768.

Raes, F., Saltelli, A., & Van Dingenen, R. (1992). Modelling formation and growth of H 2 SO 4-H 2 O aerosols: Uncertainty analysis and experimental evaluation. *Journal of Aerosol Science, 23*(7), 759-771.

Rahul, R. (2012). Kinetic modeling of pyruvate recycling pathways in pancreatic β-cells.

Rand, D. A. (2008). Mapping global sensitivity of cellular network dynamics: Sensitivity heat maps and a global summation law. *Journal of the Royal Society Interface, 5*, S59-S69.

Saltelli, A., & Bolado, R. (1998). An alternative way to compute fourier amplitude sensitivity test (FAST). *Computational Statistics & Data Analysis, 26*(4), 445-460.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global sensitivity analysis: The primer* John Wiley & Sons.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: A guide to assessing scientific models* John Wiley & Sons.

Seinfeld, J., & Lapidus, L. (1974). *Mathematical Methods in Chemical Engineering Process Modeling, Estimation, and Identification* Prentice-Hall.

Sharan, R., & Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nature Biotechnology, 24*(4), 427-433.

Sobol', I. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie, 2*(1), 112-118.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., & Doyle, J. (2004). Robustness of cellular functions. *Cell, 118*(6), 675-685.

Stephanopoulos, G., Aristidou, A. A., & Nielsen, J. (1998). *Metabolic engineering: Principles and methodologies* Academic press.

Steuer, R., & Junker, B. H. (2009). Computational models of metabolism: Stability and regulation in metabolic networks. *Advances in Chemical Physics, 142*, 105.

Van den Bos, A. (2007). *Parameter estimation for scientists and engineers* John Wiley & Sons.

Varma, A., Morbidelli, M., & Wu, H. (2005). *Parametric sensitivity in chemical systems* Cambridge University Press.

Varma, A., & Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli W3110. *Applied and Environmental Microbiology, 60*(10), 3724-3731.

Varma, A., & Palsson, B. O. (1995). Parametric sensitivity of stoichiometric flux balance models applied to wild-type escherichia-coli metabolism. *Biotechnology and Bioengineering, 45*(1), 69-79.

Walter, É., & Pronzato, L. (1990). Qualitative and quantitative experiment design for phenomenological models—a survey. *Automatica, 26*(2), 195-213.

Wang, C. (2011). Metabolic Network.

APPENDIX A

```
%% Robust Identification
function coef_min_biomass_gen_alg_20cases
A = [];
b = [];
Aeq = [];
beq = [];
 lb = [0.0135   6.75   13.5    9   0.09];
 ub = [0.0165   8.25   16.5    11  0.11];
optionsga=gaoptimset('PopulationSize',10,'Generations',10,'Display','iter');
% Fmincon
% teta =
fmincon(@for_loop_model_kd_auto_genetic_20cases,teta0,A,b,Aeq,beq,lb,ub,@nonl
inc,options);
% Genetic Algorithm
[teta,fun] =
ga(@for_loop_model_kd_auto_genetic_20cases,5,A,b,Aeq,beq,lb,ub,@nonlinc,optio
nsga);
fprintf('The best function value found was : %g\n', fun);
disp('Result teta');
disp(teta);
end



function fun=for_loop_model_kd_auto_genetic_20cases(p)
teta = p;
%% SSE calculation
theta_input=teta;
 [H]=growthModel_drv2(theta_input);
metabolites_noise_disturbances
SSE_total = 0;
for i=1:20
    SSEace=sum((Acetate(i,:)-H(1,:)).^2);
    SSEglu=sum((Glucose(i,:)-H(2,:)).^2);
    SSEoxy=sum((Oxygen(i,:)-H(3,:)).^2);
    SSEbio=sum((Biomass(i,:)-H(4,:)).^2);
    SSE_sum = SSEace + SSEglu + SSEoxy+ SSEbio;
    SSE_total = SSE_total + SSE_sum;
end
%% Sensitivity Analysis calculation
Param=[];
for ip=1:length(teta)
    par=teta(ip)*1.1;par2=teta(ip)*0.9;
    Param=[Param;par;par2];
end
km_mat = [];
dkm_mat = [];
i_teta=1;
i_teta2 = 0;
for i_param=1:length(Param)
    teta_input=teta;
    teta_input(i_teta)=Param(i_param);
    [F]=growthModel_drv1(teta_input);
            km_mat = [km_mat F];
    i_teta2 = i_teta2 + 1;
    if i_teta2 == 2
        i_teta = i_teta + 1;
        i_teta2 = 0;
    else
    end
end
%for Matrix M
Mr=[];  iteta=1; jteta=2; kteta=1;
[row,column]=size(km_mat);
for i_k=1:2:column
    input_teta=teta;
    input_param=Param;
% number 5 is from the division of 1/0.2 from the percentages of the model +
and - 10% of the value of the parameter
     ks=5.*abs(km_mat(:,iteta)-km_mat(:,jteta));
    iteta=iteta+2; jteta=jteta+2; kteta=kteta+1;
    Mr=[Mr ks];
end
% Calculation of the Matrix M for SVD
% Normalization of M to ensure that in the limit delta t --> 0 the singular
% value decomposition of M is independent of the choice of the time
% discretization delta t = ti+1 - ti
delta_t = 2;              % interval of time desired in hours
Time = 10;                % Total time of the process
M1r=(sqrt(delta_t/Time)).*Mr;
Mdot1=(sqrt(delta_t/Time)).*Mdot;
% Calculation of the Singular Value Decomposition
 [U,sigma_i,V]=svd(M1r,0);
W=transpose(V);
% Sensitivity heat map analysis
%fim = sigma_i * (max abs(Wij)) * abs(Uim(t)) Rand, 2008
maxSij=sigma_i*transpose(max(abs(W)));
fim=[];
for i_sij=1:length(maxSij)
    fi_m = maxSij(i_sij).*abs(U(:,i_sij));
    fim=[fim fi_m];
end
% lambda elimination fim
% threshold for fim is set to be 5% of the global maximum of the fim(t)
alpha=0.05*max(max(fim));
n_outputs = 5;
D2=[];
[row_f,column_f]=size(fim);
for j_f=1:1:column_f
    D1=[];
  i_step=1;j_step=n_outputs;
for i_f=1:n_outputs:row_f
    F_im=fim(i_step:j_step,j_f);
     if max(F_im)<alpha
         Fim=zeros(n_outputs,1);
     else
         Fim=F_im;
     end
     i_step=i_step+n_outputs;
     j_step=j_step+n_outputs;
     D1=[D1;Fim];
```

```
end
    D2 = [D2 D1];
end
% Calculation of Sij matrix for parameter discretization
% Sij = sigma_i * Wij
Sij=sigma_i*W;
G=(abs(Sij));
% Row elimination
 [row_g,column_g]=size(G);
uf=[]; i_g_step=1;
for i_g=1:1:row_g
if D2(:,i_g_step) == 0
    M1=zeros(1,column_g);
else
    M1=G(i_g_step,:);
end
i_g_step=i_g_step+1;
uf=[uf;M1];
end
%threshold for columns in Sij (parameter reduction)
beta=0.05*max(max(uf));
 [mpar,npar]=size(uf);
D4d=[];
for j=1:npar ;
    for i=1:mpar
    B4d=uf(i,j);
    if (B4d<beta)
        C=0;
    else
        C=B4d;
    end
    D4d=[D4d C];
    end
end
  u=reshape(D4d,[mpar,npar]);
%for a general analysis of all the metabolites and at all the times
n_outputs = 5; %number of outputs over time z(t1), z(t2), z(t3), ..., z(tn)
```

```
[row_u,column_u]=size(u);
[row_U,column_U]=size(U);
S_A_k_complete_mat = [];
for j_u=1:n_outputs:row_U
    S_A_k_mat = [];
for i_u_step=1:1:column_u
j_u_step=j_u;
 z_u_step=j_u+n_outputs-1;
 S_A_k=zeros(n_outputs,1);
for i_u=1:1:row_u
    S_A_k_vec=u(i_u,i_u_step).*abs(U(j_u_step:z_u_step,i_u));
    S_A_k = S_A_k + S_A_k_vec;
end
S_A_k_mat = [S_A_k_mat S_A_k];
%j_u_step=j_u_step+1; z_u_step=z_u_step+n_outputs;
end
S_A_k_complete_mat = [S_A_k_complete_mat; S_A_k_mat];
end
Sum_g_teta=sum(S_A_k_complete_mat);
% disp('S A K mat');
% disp(S_A_k_complete_mat);
%Percentages of the contribution of each parameter to the system
q=[];
[row_g_teta,column_g_teta]=size(Sum_g_teta);
eta=0.1;
for i_g_teta=1:1:column_g_teta
    q1=(Sum_g_teta(i_g_teta)/sum(Sum_g_teta));
    if q1<eta;
        q1=0;
    else q1=1;
    end
            q=[q q1];
end
No=sum(q);
%objective function to be maximized not considering SSE analysis
%fun= - ( ((sum(Sum_g_teta))/No));
%objective function to be minimized
```

```
% fun= ( SSE_total / ((sum(Sum_g_teta))/No));
%objective function to be maximized
% fun= - ( ((sum(Sum_g_teta))/No) / SSE_total );
end


% Diauxic Growth Model Script to Drive the Simulation in A Matlab-based Diauxic Growth Model
function [F,J]=growthModel_drv1(Param_vec);
%
global optoA optoAeq optonvar optobnd optoBeq optob optoLB optoUB optof optox constrX constrY
global constrT optooptions
global convf vglcxt vo  storeV flagSim optoexitflag optooutput optolambda KLA Km vmax Vm Va vAc teta
%
% obtain stoichiometry matrix
%
%            v1      v2     v3       v4
% Ac      -39.43       0     1.24   12.12
% Glcxt        0   - 9.46   -9.84  -19.23
% O2        - 35   -12.92   -12.73   0
% X            1    1       1      1
%
A = [-39.43,    0.0,     1.24,   12.12;
       0.0,   - 9.46,   -9.84,  -19.23;
     -35.0,   -12.92,   -12.73,    0.0;
       1.0,      1.0,    1.0,     1.0];
M     = A'*A;
[R,L] = eig(M);


optoA = A;
clear A;
%parameter assignment
convf     = [25.59331, 180.16 31.99886 60.05]; % g/mol [biomass glucose O2 acetate]
Km        = [0.015];                           % mM
KLA       = [7.5];                             % 1/h
%
```

```
% Simulation
% ----------
%
[nspecies,nreactions] = size(optoA);
%
SINIT      = [0.40 10.8 0.21 0.001];        % mM  [10.8 0.4 0.21 0.001];
%
TINIT      = 0;                            % inital time for solution
Tend       = 12.1;                           % end time for simulation
dt         = 1e-2;                         % 1e-3; %Euler step
nsteps     = ceil(Tend/dt);                % number of Euler steps
constrT    = dt;
%
s          = SINIT;
t          = TINIT;
store_s    = [];
store_t    = [];
store_dsdt = [];
store_v    = [];
%
for (istep=TINIT:dt:Tend),
    %
    [dsdt]     = growthModel_system1(t,s,Param_vec)'; % time derivatives of
states
    %
    store_s    = [store_s;s];
    store_t    = [store_t;t];
    store_dsdt = [store_dsdt;dsdt];
    store_v    = [store_v; optox'];
    %
    s       = s + dsdt*dt; % Euler integration of states
    t       = t + dt;
end;
A=interp1(store_t,store_s,[2  4  6  8  10]);
 F=reshape(A,[20,1]);
```

```
function [dy] = growthModel_system1(t,y,Param_vec);
global optoA optoAeq optonvar optobnd optoBeq optob optoLB optoUB optof optox
constrX constrY
global constrT optooptions
global convf vglcxt vo  storeV flagSim optoexitflag optooutput optolambda KLA
Km vmax Vm Va vAc teta
Ac            = y(1);
Glcxt         = y(2);
O2            = y(3);
X             = y(4);
vglcxt           = (Param_vec(4)*(Glcxt)+Param_vec(5))/(Param_vec(1)+(Glcxt));
vo               = Param_vec(3);
constrY          = y'; % stored past Y for constraint
% call to optimization routine
growthModel_opto1();
rates            = optoA*optox; % stoichiometric matrix times flux
%
dAcdt            = (X)*rates(1);
dGlcxtdt         = (X)*rates(2);
dO2dt            = (X)*rates(3) + Param_vec(2)*(0.21-O2);
dXdt             = (X)*rates(4);
%
dy = [dAcdt,dGlcxtdt,dO2dt,dXdt]';


function growthModel_opto1();
global optoA optoAeq optonvar optobnd optoBeq optob optoLB optoUB optof optox
constrX constrY
global constrT optooptions
global convf vglcxt vo  storeV flagSim optoexitflag optooutput optolambda KLA
Km vmax Vm Va vAc teta
% Solve
%    min optof'*optox
%      x
%  s.t.
%
%    M*optox <= bndM
%    optoLB<=optox<=optoUB
% objective weights
%
optof   = -1*[1 1 1 1]';            % objective function coefficients
optobnd = 100;
optob   = optobnd+zeros(4,1);
optoBeq =[];
optoLB  = -1E-6+zeros(4,1);    % lower bound on fluxes
optoUB  = 100+zeros(4,1);      % upper bound on fluxes
m1   = -1.*optoA;     % budman Yk+1 = Yk + AjXkT > 0
m2   = -1*optoA(2,:); % Glcxt uptake vglcxt = 10*Glcxt/(Km+Glcxt);
m3   = -1*optoA(3,:); % O2 uptake    vo     = 15;
 m4    = -1*optoA(1,:);       % budman       Aj     <= dY/dt upper bound
b1   = constrY./(constrY(end)*constrT);  % budman Yk+1 = Yk + AjXkT >= 0 -->
b2   = 1.0*vglcxt; % Glcxt uptake Aglucxt*v <= 10*Glcxt/(Km+Glcxt); 0.01
b3   = vo;                                % O2 uptake    AO2*v     <= 15;
 b4   = vAc;                               % budman       Aj         <= dY/dt
M    = [m1;
        m2;
        m3];
bndM = [b1;
        b2;
        b3];
optooptions = optimset('TolFun',1E-8,'MaxIter',5E6,'Display','off');
[optox,optoF,optoexitflag,optooutput,optolambda] =
linprog(optof,M,bndM,[],[], ...
optoLB,optoUB,[],optooptions);

% Set Based Constraints
function [c,ceq] = nonlinc(teta)
 [Eval]=growthModel_drv(teta);
Acet=Eval(1,:);
Gluc=Eval(2,:);
Oxyg=Eval(3,:);
Biom=Eval(4,:);
metabolites_noise_disturbances
Bio_up= max(Biomass);
Bio_low= min(Biomass);
```

```
Glu_up= max(Glucose);
Glu_low= min(Glucose);
Oxy_up= max(Oxygen);
Oxy_low= min(Oxygen);
Ace_up= max(Acetate);
Ace_low= min(Acetate);

c(1) = Acet(1,5) - Ace_up(1,5);
c(2) = -Acet(1,5) + Ace_low(1,5);
c(3) = Acet(1,9) - Ace_up(1,9);
c(4) = -Acet(1,9) + Ace_low(1,9);
c(5) = Acet(1,13) - Ace_up(1,13);
c(6) = -Acet(1,13) + Ace_low(1,13);
c(7) = Acet(1,17) - Ace_up(1,17);
c(8) = -Acet(1,17) + Ace_low(1,17);
c(9) = Acet(1,21) - Ace_up(1,21);
c(10) = -Acet(1,21) + Ace_low(1,21);

c(11) = Gluc(1,5) - Glu_up(1,5);
c(12) = -Gluc(1,5) + Glu_low(1,5);
c(13) = Gluc(1,9) - Glu_up(1,9);
c(14) = -Gluc(1,9) + Glu_low(1,9);
c(15) = Gluc(1,13) - Glu_up(1,13);
c(16) = -Gluc(1,13) + Glu_low(1,13);
c(17) = Gluc(1,17) - Glu_up(1,17);
c(18) = -Gluc(1,17) + Glu_low(1,17);
c(19) = Gluc(1,21) - Glu_up(1,21);
c(20) = -Gluc(1,21) + Glu_low(1,21);

c(21) = Oxyg(1,5) - Oxy_up(1,5);
c(22) = -Oxyg(1,5) + Oxy_low(1,5);
c(23) = Oxyg(1,9) - Oxy_up(1,9);
c(24) = -Oxyg(1,9) + Oxy_low(1,9);
c(25) = Oxyg(1,13) - Oxy_up(1,13);
c(26) = -Oxyg(1,13) + Oxy_low(1,13);
c(27) = Oxyg(1,17) - Oxy_up(1,17);
c(28) = -Oxyg(1,17) + Oxy_low(1,17);
```

```
c(29) = Oxyg(1,21) - Oxy_up(1,21);
c(30) = -Oxyg(1,21) + Oxy_low(1,21);

c(31) = Biom(1,5) - Bio_up(1,5);
c(32) = -Biom(1,5) + Bio_low(1,5);
c(33) = Biom(1,9) - Bio_up(1,9);
c(34) = -Biom(1,9) + Bio_low(1,9);
c(35) = Biom(1,13) - Bio_up(1,13);
c(36) = -Biom(1,13) + Bio_low(1,13);
c(37) = Biom(1,17) - Bio_up(1,17);
c(38) = -Biom(1,17) + Bio_low(1,17);
c(39) = Biom(1,21) - Bio_up(1,21);
c(40) = -Biom(1,21) + Bio_low(1,21);
ceq = [];
end
```