

**Parameter Estimation
in Nonlinear Dynamical Systems**

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus,
Prof.dr. J.J.M. Franse
ten overstaan van een door
het college van promoties ingestelde commissie
in het openbaar te verdedigen in de Aula der Universiteit
op donderdag 12 maart 1998 te 11.00 uur

door

Walter Johannes Henricus Stortelder
geboren te Lichtenvoorde

CWI
Amsterdam

Parameterschatten in niet-lineaire dynamische systemen.

Promotor: Prof.dr. P.W. Hemker

Co-promotor: Prof.dr. C.A.J. Klaassen

Faculteit: Wiskunde, Informatica, Natuurkunde en Sterrenkunde

Het hier beschreven onderzoek is mogelijk gemaakt door financiële steun van de Technologiestichting (STW) onder projectnummer: CWI22.2695.

This one goes out to the ones I love
(Freely rendered from R.E.M.)

Preface

This thesis contains the results of my PhD project at CWI in Amsterdam. Of course, the work involved is not completely mine. During this period I had the pleasure and privilege to learn, enjoy and grow. In this aspect the guidance and patience of Piet Hemker was not a negligible parameter but a big constant. He showed me how to distinguish between essence and details, to be efficient and not to waste time on pursuing dead-ends.

In a later stage of the project Chris Klaassen, who could combine hospitality with a busy agenda, got involved. Even so, with every succeeding appointment we made, my fear increased for the consequences to his crowded schedule.

During the first year of my term I collaborated with Kees Everaars on the implementation of the various versions of the software. The multi-topic talks in between were warm and frequent.

I'm grateful to the people at CWI who create a perfect environment to get spoilt in many aspects. With respect to the MAS-sers, I would like to thank them all for the various occasions when we met. I would like to mention two of them explicitly. First, my room mate Jaap Noordmans, I'm not only impressed by your fast biking and good morning coffee. Second, Jacques de Swart, with whom it is nice to disagree. You know how to motivate people and talk them into things they might regret afterwards, or at least would not have done without your persuasion.

To the Dutch Technology Foundation (STW) I'm not only indebted for the financial support, but also for their formation of a users' committee. This committee gave me more a live line than a dead end. The input I got during the vivid meetings kept me in touch with the applications, although each member wanted more than I could offer.

Special thanks go to Peter Verheijen, who spent his holiday with a preprint of my thesis on his lap and returned with a pre-preprint. Your remarks were valuable, and your criticism and deep interest kept me alert. Another person who contributed by careful reading is Eligius Hendrix.

During 'field work' periods, I had the opportunity to look into the kitchen of Coen Hemker at the biochemistry department, University Maastricht, and of Rein van der Hout at Akzo-Nobel in Arnhem. Both visits resulted in substantial contributions to Chapter 6. Another kitchen with a different taste was opened by Jose Merchuk. Although I'm still not sure whether he invited me to explore the country and its history, or to produce a scientific paper, I'm sure he managed to do both.

Paul de Bruin not only inspired me to write another section in Chapter 6, but also gave me the permanent frustration of never winning in badminton or snooker. Another

article, which is not a part of this thesis, but related to this research, was written with János Pintér and Jacques de Swart.

I gathered and stayed in touch with many other friends and, although writing names is dangerous, I shall take the risk here by stating that some of them are more special than others: Alien, Frank, Gerard, Jet, John (2×), Leon, Marc, Paul and Rob (in alphabetical order). My house mates gave me the opportunity to take my mind off work during numerous dinners and to combine daily routine with pleasure.

Probably I kept puzzling my family when I explained them what I was doing in the ‘big city’. The main thing they understood was that it needs a lot of patience and support. This is exactly what I received from all of them in the pursuit of my task over the years. Thanks to all of you.

Last but by no means least, I owe a lot to a person who managed to encourage me during rainy days. Iris, I looked into my directories and found out that your mails take more disk space than the text files of this thesis. Your talk requests, which were always irrisistable, will soon be I2I. Toda raba is an understatement here and I’m sure you are not able to estimate its extent and depth in this context.

Walter Stortelder, A~~l~~dam, December 1997.

Contents

Preface	i
Contents	iii
Nomenclature	vii
1 Introduction and Outline	1
1.1 Mathematical formulation	4
1.2 Fitness criterion	5
1.3 Variational equations	5
1.4 Numerical solution of the model equations	7
1.5 Minimisation	8
1.6 Statistical background	9
1.7 Parameter constraints	12
1.8 A case study from biochemistry	13
1.9 A case study from population dynamics	15
1.10 Concluding remarks	16
Appendix 1.A	16
Appendix 1.B	16
Appendix 1.C	16
2 Parameter Estimation by Total Least Squares	21
2.1 Introduction	21
2.2 Mathematical description of TLS	22
2.3 Statistical background	25
2.4 Total least squares with parameter constraints	26
2.5 Conclusions	28
3 Maximum Likelihood Estimators	29
3.1 Introduction	29
3.2 Least squares criterion	29
3.3 Weighted least squares criterion	31
3.3.1 A priori known weights	31
3.3.2 Unknown weights	31

3.4	Numerical computation (independent case)	35
3.5	Dependent measurement errors	37
3.6	Numerical computation (dependent case)	39
3.7	MLE and total least squares	41
3.7.1	A priori known weights	41
3.7.2	Unknown weights (TLS)	42
3.7.3	Independent measurement errors	42
3.7.4	Dependent measurement errors	43
3.8	L_1 -optimisation and Laplace distribution	44
3.9	Conclusions	45
4	Nonlinear Regression, Bias and Curvature	47
4.1	Overview of the chapter	47
4.2	Linear Regression	48
4.3	Biased estimators	50
4.3.1	Analytic result	50
4.3.2	Monte Carlo	52
4.3.3	Bias measure of Box	53
4.4	Curvature measures	54
4.5	Investigation of levelsets	61
4.6	Parameter constraints and redundancy	63
4.7	Concluding remarks	64
5	Optimal Experiment Design	65
5.1	Introduction	65
5.2	Problem formulation	65
5.3	Parameter - state variable dependence	67
5.4	OED and improved confidence regions	70
5.4.1	Design criteria	70
5.4.2	Repeated design	71
5.4.3	Sequential design	73
5.5	OED and model discrimination	76
5.6	OED and nonlinearity	77
5.7	Concluding remarks	77
6	Case Studies	79
6.1	Production of resins	79
6.1.1	Introduction	79
6.1.2	Reaction mechanism	80
6.1.3	Experiments performed	82
6.1.4	Model equations	82
6.1.5	Treatment of the melamine concentrations	83

6.1.6	Parameter estimation	84
6.1.7	Reparametrisation and results	84
6.1.8	Conclusions	88
6.2	Mathematical modelling in blood coagulation	92
6.2.1	Introduction	92
6.2.2	Experimental data	93
6.2.3	Reaction mechanism	94
6.2.4	Model equations	95
6.2.5	Parameter estimation and model validation	98
6.2.6	Results	100
6.2.7	Conclusions	102
6.3	Production by plant cells in suspension	104
6.3.1	Introduction	104
6.3.2	Materials and methods	105
6.3.3	Model development	106
6.3.4	Experimental results and final model confirmation	109
6.3.5	Conclusions	112
6.4	ZLA-kinetics	117
6.4.1	Description of the chemical reactions	117
6.4.2	Problem description of ZLA-kinetics	118
6.4.3	Parameter estimation results for ZLA-kinetics	119
6.4.4	Conclusions and remarks for further research	120
6.5	Water penetration in an aramide yarn	122
6.5.1	Introduction to the problem	122
6.5.2	Proposed models	122
6.5.3	Numerical implementation	124
6.5.4	Model discrimination and parameter estimation	124
6.5.5	Conclusions	126
6.6	Macroeconomic time series	128
6.6.1	Introduction	128
6.6.2	Derivation of candidate models	128
6.6.3	Comparison of prediction results	130
6.6.4	Concluding remarks	131
6.7	The DOW problem	132
6.7.1	Introduction	132
6.7.2	Description of the problem	132
6.7.3	Data	134
6.7.4	Results	134
6.7.5	More general model equations	138
6.7.6	Concluding remarks	140
Appendix 6.A	143
Appendix 6.B	147

7 Software Design and Implementation	149
7.1 Introduction	149
7.2 Design principles of the application spIds	150
7.3 Structure of the software	152
7.4 The modelfile	153
7.5 The datafile	157
7.6 Algebraic Engine	160
7.7 Filter	161
7.8 Numerical engine	161
7.9 Graphical user interface (GUI)	162
7.10 Database manager	163
7.11 Concluding remarks	164
Samenvatting	165
Bibliography	167
Index	173

Nomenclature

<u>symbol</u>	<u>meaning</u>	<u>dimension</u>
α	probability of excess	1
α_i	pre-exponential factor	1
Γ^\perp	relative intrinsic curvature	1
Γ^\parallel	relative parameter-effect curvature	1
Δ^D	dependent confidence region	1
Δ^I	independent confidence region	1
ε	vector of measurement errors	N
θ	vector of unknown parameters or regression variables	m
θ^*	true parameter vector	m
$\hat{\theta}$	estimated parameter vector	m
λ_i	i -th singular value	1
σ, σ^2	standard deviation and variance, respectively, of a normal distribution	
τ	discrepancies related to the independent variable	N
χ_i^2	Chi-square distribution with i degrees of freedom	
c_i	component of i -th measurement ($1 \leq c_i \leq n$)	1
$d(\theta)$	(unweighted) discrepancy vector	N
$D(\theta)$	matrix with unweighted discrepancies	$r \times q$
\mathbf{E}	expectation	
E_i	activation energy	1
$\mathcal{F}_\alpha(i, j)$	upper α quantile for Fisher's F-distribution with i and j degrees of freedom	1
H	Hessian ¹ matrix	$N \times m \times m$
J	Jacobian matrix	$N \times m$
J_{add}	Jacobian matrix due to additional measurements	$N_{add} \times m$
K	number of constraints with respect to the parameters, θ	1

¹Many people write Hessian, which does not seem to honour the German mathematician Ludwig Otto Hesse (1811-1874) and it is also not consistent (compare e.g. Boolean).

<u>symbol</u>	<u>meaning</u>	<u>dimension</u>
l	number of independent variables	1
m	number of unknown parameters	1
M	moment matrix	$q \times q$ ($2q \times 2q$)
n	number of dependent variables	1
N	number of measurements	1
$\mathcal{N}(\mu, V)$	Gaussian or normal distribution with mean μ and covariance matrix V	
N_{add}	number of additional measurements	1
N_{MC}	number of Monte Carlo (MC) simulations	1
p	probability	1
q	number of measured components ($q \leq n$)	1
r	number of samples ($qr \geq N$)	1
$R(\theta)$	(nonlinear) constraints on the parameters	K
s	estimator or estimate of σ	1
t	independent variable (for the special case that $l = 1$ x is replaced by t)	1
V	covariance matrix of the measurement errors	$N \times N$, $q \times q$ or $2q \times 2q$
w_i	weight corresponding to the i -th measurement	1
x	vector of independent, regression or explanatory variables	l
y	vector of dependent or response variables	n
y'	derivative of y with respect to t	n
\tilde{y}_i	i -th measured value	1
Y	vector of (weighted) discrepancies ²	N

²It should be emphasised here that the discrepancies, $d_i(\theta) = y_{c_i}(x_i, \theta) - \tilde{y}_i$ ($i = 1 \dots, N$), depend on the parameters and that the residuals are the discrepancies evaluated for the estimated values of the parameters, $d(\hat{\theta})$.

Chapter 1

Introduction and Outline

Many processes from (bio-)chemistry, geo-sciences, biology, electrical and mechanical engineering or econometrics can be mathematically described by systems of differential algebraic equations (DAEs). These equations describe the dynamical behaviour of the processes under consideration. For example, in the case of a chemical reaction, concentrations change in time due to chemical interactions between the substances involved. Then the independent variable is time. If the concentrations are not constant over the reactor, we have additional space coordinates as independent variables and end up with partial differential equations (PDEs). This thesis will only consider systems with one independent variable, except for a single example where a problem described by PDEs is reduced to a system of DAEs. The dependent variables –still considering a chemical reaction– correspond to the concentrations of the chemical substances of interest during the reaction. Starting from a given initial situation, i.e. known values of the state variables at a given initial time, the reaction begins. The solution of the model equations gives an approximation for the concentrations of the substances in time. With the exception of class room examples, models from real-life applications yield equations which have to be solved with the use of dedicated numerical software, often in combination with powerful hardware.

The model of interest, the one which gives a satisfactory description of the process under consideration, is nearly always the result after a period with intensive and extensive communication between the experimentalist and the modeller. The evolution of a model takes time and asks for skills and experience of both the experimentalists and the mathematical modellers. The research in this thesis not only focuses on mathematical tools which make the process of modelling less time consuming and more transparent, but –in a number of cases– it also reflects this process of interaction between experimentalist and mathematician. Further, it deals with the software aspects and actual implementation of a computer program which enables the experimentalist to investigate the mathematical model easily.

A mathematical model, or a set of candidate models is based on experience and physical insight from the application domain. The model equations are set up in such a way that their outcome is in accordance with well established facts of the physical process studied. In order to validate models, to discriminate between models or to calibrate

models we need to compare the model outcome with measurements. We suppose that the final model gives a sufficiently precise description of the process under consideration, which implies that no model errors are present or that they can be neglected. Of course, for practical problems this seems an ideal situation and one might think it is a naive approach, but the absence of better alternatives and the valuable results in many real-life cases justifies this method. We do not believe there is a ‘true’ model, but we assume that it is possible to study a model whose model errors are an order of magnitude smaller than the uncertainties in the measurements.

In the case of parameter estimation or model calibration, we calculate the best fitting model from a continuum of models. We consider models which are expressed mathematically by systems of differential algebraic equations (DAEs) with a certain degree of freedom, expressed by the presence of a set of parameters. If we return to the example from chemistry, these parameters may correspond with unknown reaction rates or unknown initial concentrations, which cannot be obtained by means of direct observation or from other resources. These unknown parameters are computed such that the discrepancies between the theoretical model output and the measured data are minimal in some sense: the calculated, theoretical values or model responses should fit the measurements. The choice for a certain fitness criterion depends on the knowledge and the assumptions about statistical properties of the measurement errors. After fitting the model to the data, not only the final estimates of the unknown parameters are of interest, but also information about their reliability. When we adjust parameters we have –strictly speaking– a different model, but we will not make this distinction throughout this thesis. We consider two models, M_1 and M_2 , with their corresponding vectors of unknown parameters θ and ϕ , respectively, to be the same, if for every choice of θ there exists exactly one ϕ such that the models $M_1(\theta)$ and $M_2(\phi)$ yield the same model responses. This means that $\dim(\theta) = \dim(\phi)$ and that reparametrisation via a bijective mapping does not change the model, although it may have other consequences, e.g., for the nonlinearity of the parameter estimation problem.

Model validation, model reduction and model selection is a systematic process that eventually leads to the recommendation of one model or a set of models that is (i) consistent with the data, (ii) in accordance with well established facts concerning the physical process and (iii) not unnecessarily complex. In each step of the process the lack of fit is expressed in statistical quantities on the basis of which we accept or reject a model, simplify it or choose between models. This process is closely related to the design of experiments. If on the basis of the available data no decisive answers with respect to model selection can be derived, advice with respect to a setup for additional experiments is needed.

Much research has already been carried out to estimate unknown parameters by fitting a numerical solution to a set of experimental data. Many publications consider the case where the model response can be obtained relatively easily: the state variables can be written explicitly as function of the independent variables and the parameters [Bar74, Rat83, SW88, BW88]. In these references, the emphasis is rather on the theoretical and

statistical aspects than on practical implementation and numerical aspects of nonlinear regression with models given by DAEs. The latter case is considered in more detail in [Hem72a, Boc85, Sch85]. Although some authors might give slightly different definitions, parameter estimation in dynamical systems is essentially the same as nonlinear regression where the model is given by a set of DAEs. In current literature it seems that there is still a gap between numerical mathematics and nonlinear regression analysis. In this thesis I try to fill this gap partially by merging ideas from the whole spectrum of tools and ideas involved in the broad field of parameter estimation in nonlinear dynamical systems with an accent on normal measurement errors.

In this first chapter we start with a mathematical formulation of our parameter estimation problems in DAEs. The measurement errors are assumed to be normally distributed, stochastically independent and only present in the dependent variables. The variances of the measurement errors are known, or known up to a constant of proportionality. Based on fundamental statistics and under these conditions we have to minimise the sum of the squared discrepancies between model responses and measurements. This approach is known as *Ordinary Least Squares (OLS)* estimation. Thereupon, in this chapter we present numerical techniques to solve this problem. Two additional sections about the statistical background and constraints on the parameters are followed by two introductory case studies from biochemistry and population dynamics.

In Chapter 2 we deal with total least squares (TLS), where the structure of the chapter is analogous to the structure of the present chapter on OLS. In the case the measurement errors with respect to the independent variables are zero or negligible, OLS approaches can be used. If this is not the case, the TLS approach should be applied instead. The extensions from OLS to TLS are described in Chapter 2. A stable and efficient algorithm to deal with TLS estimation is presented, in combination with an overview of the additional consequences concerning the statistical background and parameter constraints.

In the case the measurement errors are known to have a normal distribution, but their variances –and therefore the weights– are not known a priori, these quantities can be estimated together with the parameters if a few assumptions are made. This case with unknown weights, both for OLS and TLS, with independent and dependent measurement errors is dealt with in Chapter 3. In this chapter we also introduce an algorithm to compute L^1 -estimates, when the sum of the absolute discrepancies has to be minimised. This approach is used if the measurement errors come from a Laplace distribution. It is known to be more robust –i.e. less sensitive to outliers– than least squares methods. This characteristic makes it attractive in combination with a least squares approach, as a kind of two-stage method, when no good initial estimates for the parameters are available. The first guess for the parameters is improved by minimising the sum of absolute discrepancies, subsequently the resulting L^1 -estimate is used as an initial parameter guess for least squares estimation.

It should be emphasised that the statistical results concerning the confidence regions for OLS and TLS estimates are obtained by linearisation. For most nonlinear problems

this gives quite accurate information in a sufficiently small neighbourhood of the minimum. But if we restrict ourselves to this information from the linearisation it can be very misleading for strongly nonlinear problems. It may turn out that the confidence region of interest is no ellipsoid at all –as follows from linear theory– but a non-convex and irregular region. Therefore we have to verify how accurate the linear approximation is. More information concerning nonlinearity, bias of the estimates, curvature and related topics is found in Chapter 4.

Chapter 5 deals with optimal experiment design. Given a model, a set of estimated parameters and the corresponding confidence regions, it deals with the question which additional measurements should be performed to increase the reliability of the parameter estimates. Or, given two models, which measurements should be performed to be able to discriminate between the two models.

A variety of case studies, from (bio-) chemistry, physics, econometrics, is described in Chapter 6. All case studies have been carried out in collaboration with researchers from other disciplines. In most cases they supplied the data and a number of possible models. After receiving the first candidate model(s) and the data, usually numerous improvements have been made regarding many aspects of modelling in order to come up with an appropriate model.

The setup and the implementation of the software used for the computations is described in Chapter 7. The chapter starts with a description of the way problem-dependent input is specified: the format for the mathematical model and the experimental data. The software contains computer algebra routines for automatic generation of model dependent program parts and numerical routines for the solution of the differential algebraic equations, minimisation of the fitness criterion and statistical analysis of the computed estimates. The DAE solver is geared to solve these model equations in combination with the sensitivity equations. A graphical user interface (GUI) has been developed to steer through the computation in order to influence the precise formulation of the parameter estimation problem during the calculation, and to view the numerical results by direct visualisation.

1.1 Mathematical formulation

The model equations are given by the system of *differential algebraic equations (DAEs)*,

$$Ay' = A \frac{dy}{dt} = f(t, y, \theta), \quad \text{with } y(t_0, \theta) = y_0(\theta), \quad (1.1)$$

where t denotes time, θ is an m -dimensional vector of unknown *parameters*, $y(t, \theta)$ is an n -dimensional *state vector* depending on t and θ , the function $f(t, y, \theta)$ maps $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m$ into \mathbb{R}^n and A is a constant $n \times n$ -matrix. In the simplest case A is a diagonal matrix with $A_{ii} = 1$ if the i -th equation is a differential equation and $A_{ii} = 0$ if the i -th equation is algebraic.

In order to estimate the unknown parameters, a number of measurements, say N , are available for the process under consideration. Each *measurement* is characterised by the triple

$$(c_i, t_i, \tilde{y}_i) , \quad i = 1, \dots, N , \quad (1.2)$$

where c_i indicates which *component* of the state vector, y , has been measured, t_i is the time of the measurement and \tilde{y}_i is the measured value. Of course, a necessary condition to estimate the unknown parameters is that the number of parameters, m , does not exceed the number of measurements, N , i.e. $m \leq N$. The solution of (1.1) for the c_i -th component at time t_i , which corresponds to the i -th measurement, is denoted by $y_{c_i}(t_i, \theta)$.

1.2 Fitness criterion

The fitness criterion depends on the discrepancies between the calculated and the measured values. The vector of discrepancies reads:

$$d(\theta) = (y_{c_i}(t_i, \theta) - \tilde{y}_i)_{i=1, \dots, N} . \quad (1.3)$$

A usual approach is to estimate the unknown parameters such that the (weighted) sum of squared discrepancies:

$$S(\theta) = \sum_{i=1}^N w_i^2 d_i^2(\theta) , \quad (1.4)$$

is minimal. The positive weights, w_i , are based on the accuracy of the measurements and have dimension $1/[\tilde{y}_i]$. In the case the errors in the measurements are stochastically independent and normally distributed with standard deviation σ_i , and if we take w_i proportional to $1/\sigma_i$, weighted least squares yields the maximum likelihood estimate (MLE). The value of θ which minimises (1.4) is called the weighted least squares estimate and is denoted by: $\hat{\theta}$. Summarising: In the case of (i) normally distributed and independent measurement errors, (ii) the above choice of the weights, and (iii) a negligible measurement error for the independent variable, t , minimisation of (1.4) leads to the most likely value for the parameter vector, $\hat{\theta}$. This is discussed in more detail in Chapter 3.

1.3 Variational equations

In order to use a *gradient-based minimisation* procedure and to perform a statistical analysis we solve, besides the set of DAEs (1.1), also the corresponding set of variational or *sensitivity equations* with respect to the unknown parameter vector. This leads to an additional set of nm DAEs, written in a compact matrix notation as:

$$A \frac{d}{dt} \frac{\partial y}{\partial \theta} = \frac{\partial f}{\partial \theta} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \theta} , \quad \text{with} \quad \frac{\partial y(t_0, \theta)}{\partial \theta} = \frac{\partial y_0(\theta)}{\partial \theta} . \quad (1.5)$$

The solution of (1.5) yields the gradient $\partial y(t, \theta)/\partial \theta$, which will be used for the minimisation of the weighted sum of squared discrepancies (cf. (1.4)) and the statistical analysis in Section 1.6.

If we write down (1.5) explicitly and add (1.1), we obtain the complete system of equations to be solved:

$$\begin{aligned} Ay' &= f(t, y, \theta), & y(t_0, \theta) &= y_0(\theta), \\ A \frac{\partial y'}{\partial \theta_1} &= \frac{\partial f}{\partial \theta_1} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \theta_1}, & \frac{\partial y(t_0, \theta)}{\partial \theta_1} &= \frac{\partial y_0(\theta)}{\partial \theta_1}, \\ &\vdots & &\vdots \\ A \frac{\partial y'}{\partial \theta_m} &= \frac{\partial f}{\partial \theta_m} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \theta_m}, & \frac{\partial y(t_0, \theta)}{\partial \theta_m} &= \frac{\partial y_0(\theta)}{\partial \theta_m}. \end{aligned} \quad (1.6)$$

The system of equations (1.6) contains one subsystem of n nonlinear DAEs and m subsystems of the same size, which depend nonlinearly on y and linearly on $\partial y/\partial \theta_i$. The Jacobian of the overall system reads¹:

$$Jac = \begin{pmatrix} \frac{\partial f}{\partial y} & 0 & \dots & 0 \\ \frac{\partial^2 f}{\partial \theta_1 \partial y} + \frac{\partial^2 f}{\partial y^2} \frac{\partial y}{\partial \theta_1} & \frac{\partial f}{\partial \theta_1} & 0 \dots & 0 \\ \vdots & 0 & \ddots & 0 \\ \frac{\partial^2 f}{\partial \theta_m \partial y} + \frac{\partial^2 f}{\partial y^2} \frac{\partial y}{\partial \theta_m} & 0 & \dots & \frac{\partial f}{\partial \theta_m} \end{pmatrix}, \quad (1.7)$$

Inspection of (1.7) shows the one-way coupling of the system. Using a BDF method to solve the (possibly stiff) system, we can take advantage of this structure by first calculating y at each step of the numerical integration and subsequently all $\partial y/\partial \theta_i$. We also see that the Jacobian matrix of the overall system has the same eigenvalues as $\partial f/\partial y$, which is the Jacobian of the model equations. This means that the variational equations inherit the stiffness character of the original equations.

For purposes which will become clear in Chapter 4, we sometimes need second order derivatives of the state variables with respect to the parameters. This leads to an additional set of nm^2 DAEs, which can be derived by differentiation of (1.5):

$$\begin{aligned} A \frac{d}{dt} \frac{\partial^2 y}{\partial \theta^2} &= \frac{\partial^2 f}{\partial \theta^2} + 2 \frac{\partial^2 f}{\partial \theta \partial y} \frac{\partial y}{\partial \theta} + \frac{\partial^2 f}{\partial y^2} \left(\frac{\partial y}{\partial \theta} \right)^2 + \frac{\partial f}{\partial y} \frac{\partial^2 y}{\partial \theta^2}, \\ \text{with} \quad \frac{\partial^2 y(t_0, \theta)}{\partial \theta^2} &= \frac{\partial^2 y_0(\theta)}{\partial \theta^2}. \end{aligned} \quad (1.8)$$

¹To be more precise we need a second Jacobian; the derivative of (1.6) with respect to $(y', \partial y'/\partial \theta_1, \dots, \partial y'/\partial \theta_m)$, which equals $I_m \otimes A$. This second Jacobian is taken care of in the numerical solver and does not influence the inheritance of the stiffness character.

The solution of (1.8) corresponds to second order information which can be used to investigate the nonlinearity of a parameter estimation problem. It can also be used for the minimisation of the residual sum of squares by Newton's method, as will be shown at the end of Section 1.5. Analogously to the derivation of the Jacobian in (1.6), we can derive the Jacobian of (1.8). We omit this exercise here. Relevant is that it shows that also (1.8) inherits the stiffness character of (1.1).

1.4 Numerical solution of the model equations

In this section we assume the reader to be familiar with the theory of differential algebraic equations and their numerical solution. For the other sections a basic understanding of the solution method for the model equations is not necessary and it can be regarded as a black box which produces the values $y_{c_i}(t_i, \theta)$ and the corresponding derivatives $\partial y_{c_i}(t_i, \theta)/\partial \theta$, and –if required– $\partial^2 y_{c_i}(t_i, \theta)/\partial \theta^2$. For the actual implementation, knowledge of the numerical solution method and the stiffness behaviour of the sensitivity equations is required in order to transform an existing DAE solver into a special purpose solver for (1.6). An introduction to differential algebraic equations and their numerical solution can be found in, e.g., [Gea71, HNW93, HW96].

In the case of differential equations, $A = I_n$, in general the model equations (1.1) are stiff. This is due to the presence of fast and slow phenomena in the processes they originate from. For the differential algebraic equations, we restrict ourselves to systems of index 1 only. In both cases the equations have to be solved by an implicit method.

The fact that the size of the problems we encountered was relatively small, $n < 100$, and the possibility to solve the variational equations by making full use of the same stiffness character, made us decide to choose a numerical solution method based on the *backward differentiation formulae* (BDF). If a proper BDF method, with a certain order and step-size strategy is provided to solve (1.1) numerically, the same strategy can be used to integrate (1.5) and (1.8) numerically.

When parameter estimation is put into practice the choice of an efficient solver for the model and variational equations is of major interest because more than 80% of the computation time is used for the integration of these equations.

The use of the variational equations in combination with the same order and step strategy, leads to a faster and more accurate gradient than is possible by finite differences. In practice, generating analytic, derivative functions is no impediment as it can be done automatically by a *computer algebra* package (we use *MAPLE V*, see [CGG⁺91]).

A third alternative to retrieve derivatives is proposed in [BCC⁺92]. This approach is based on automatic differentiation and deserves further investigation in this context. We did not consider this method in this study.

1.5 Minimisation

Introducing the vector of weighted discrepancies as the column vector

$$Y(\theta) = (w_i d_i(\theta))_{i=1, \dots, N} , \quad (1.9)$$

we write the sum of squares (1.4) as

$$S(\theta) = \|Y(\theta)\|^2 = Y^T(\theta)Y(\theta) . \quad (1.10)$$

For a given value of θ , the vector $Y(\theta)$ can be computed by numerical integration of (1.1). The variational equations (1.5) facilitate the calculation of the $N \times m$, Jacobian matrix

$$J(\theta) = \frac{\partial Y(\theta)}{\partial \theta} = \left(w_i \frac{\partial}{\partial \theta} y_{c_i}(t_i, \theta) \right)_{i=1, \dots, N} . \quad (1.11)$$

Minimisation of (1.10) is done by an iterative procedure. Suppose θ is a trial vector and its correction is given by $\delta\theta$. The squared sum of the improved parameter vector can be approximated by a quadratic function of $\delta\theta$

$$\begin{aligned} S(\theta + \delta\theta) &= Y^T(\theta + \delta\theta)Y(\theta + \delta\theta) \\ &\approx (Y(\theta) + J(\theta)\delta\theta)^T(Y(\theta) + J(\theta)\delta\theta) \\ &= Y^T(\theta)Y(\theta) + 2\delta\theta^T J^T(\theta)Y(\theta) + \delta\theta^T J^T(\theta)J(\theta)\delta\theta . \end{aligned} \quad (1.12)$$

The minimum of the quadratic form is given by the *normal equations*:

$$J^T(\theta)J(\theta)\delta\theta = -J^T(\theta)Y(\theta) . \quad (1.13)$$

This formula is the starting point for a *Gauss-Newton* type method. From (1.13) it is clear that the Gauss-Newton procedure fails if the matrix $J(\theta)$ is (almost) singular. A well known remedy is the use of the *Levenberg-Marquardt* method to stabilise the procedure [Mar63, DS83]. This method replaces (1.13) by

$$(J^T(\theta)J(\theta) + \lambda I_m)\delta\theta = -J^T(\theta)Y(\theta) , \quad (1.14)$$

where λ is adjusted on the basis of the condition number of the matrix $J(\theta)$. The Levenberg-Marquardt method can be seen as a hybrid method between Gauss-Newton and *steepest descent*.

To solve $\delta\theta$ from (1.14), we use the *singular value decomposition (SVD)* of the matrix $J(\theta)$ defined by

$$J(\theta) = U(\theta)\Sigma(\theta)V^T(\theta) , \quad (1.15)$$

where $U(\theta)$ and $V(\theta)$ are $N \times m$ and $m \times m$ unitarian matrices, respectively, such that $U^T(\theta)U(\theta) = I_m$ and $V^T(\theta)V(\theta) = V(\theta)V^T(\theta) = I_m$. The $m \times m$ -matrix $\Sigma(\theta)$ is diagonal

and contains the singular values in a non-increasing order [GV83]. Substitution of (1.15) in (1.14) leads to the following expression for the correction of the parameter vector

$$\delta\theta = -V(\theta) (\Sigma^2(\theta) + \lambda I_m)^{-1} \Sigma(\theta) U^T(\theta) Y(\theta) . \quad (1.16)$$

Upon convergence of the Levenberg-Marquardt algorithm we obtain a final or least squares estimate of θ , denoted by $\hat{\theta}$.

Another possibility to minimise $S(\theta)$ is by *Newton's method*, which needs second order derivatives. Therefore, we introduce the $N \times m \times m$, Hessian matrix :

$$H_{ijk}(\theta) = w_i \frac{\partial^2 y_{c_i}(t, \theta)}{\partial \theta_j \partial \theta_k} . \quad (1.17)$$

Now, instead of the expansion (1.12), we write:

$$\begin{aligned} S(\theta + \delta\theta) \approx & Y^T(\theta)Y(\theta) + 2\delta\theta^T J^T(\theta)Y(\theta) + \\ & \delta\theta^T (J^T(\theta)J(\theta) + Y^T(\theta)H(\theta)) \delta\theta . \end{aligned} \quad (1.18)$$

Deriving $\delta\theta$ from this last expression leads to Newton's method, where the Gauss-Newton method and its variants neglect the additional term $Y^T(\theta)H(\theta)$ [DS83]. Although the Hessian can be computed via the same order and step strategy as explained in Section 1.4, to our experience the additional computational time does not result in faster or more accurate final estimates. Therefore, we stick to the Levenberg-Marquardt method and only use the Hessian in order to perform local analyses in the vicinity of $\hat{\theta}$.

1.6 Statistical background

Let the measurement error of the i -th measurement be denoted by ε_i . We assume that there exists a model which is close enough to reality such that for the 'true' parameter vector, θ^* , the equation

$$\tilde{y}_i = y_{c_i}(t_i, \theta^*) + \varepsilon_i$$

is valid or at least gives a close approximation and expresses a reasonable and workable assumption. In this section the errors in the measurements are considered (i) to be normally distributed, (ii) to have zero expectation and (iii) to be stochastically independent. The measurement errors are scaled by their weights such that they get a constant variance², σ^2 . Notice that this setup of scaling the measurement errors can be applied both for absolute and relative measurement errors. The (weighted) experimental errors in the measurements are given then by $Y(\theta^*)$, as in (1.9). This implies that the covariance matrix of the experimental errors is given by:

$$\mathbf{E} (Y(\theta^*)Y^T(\theta^*)) = \sigma^2 I_N . \quad (1.19)$$

²In general the separate standard deviations, σ_i , are approximately known up to an unknown factor of proportionality. This factor, denoted by $1/\sigma$, can be estimated after the optimal estimates of the parameters have been calculated.

We assume that the matrix $J(\hat{\theta})$ from (1.11) is regular³. Further, we notice that the additional term λI_m in (1.16) is introduced only for stabilisation of the numerical minimisation problem; it has no influence on the solution found and it does not play a role in the statistical analysis. As a consequence of these remarks and (1.19), we may approximate the covariance matrix of $\Delta\theta = \theta^* - \hat{\theta}$ by⁴:

$$\mathbf{E}(\Delta\theta\Delta\theta^T) \approx \sigma^2 \left(J^T(\hat{\theta})J(\hat{\theta}) \right)^{-1} = \sigma^2 V(\hat{\theta})\Sigma^{-2}(\hat{\theta})V^T(\hat{\theta}) , \quad (1.20)$$

which is a linear approximation. Within the order of this approximation, the unknown $J(\theta^*)$ can be replaced by $J(\hat{\theta})$ under very general conditions, as derived in [SW88, Section 2.1.2]. *All statements below hold exactly if the discrepancies, $d_i(\theta)$, are linear in θ , but in the more general case we consider, they hold approximately only.* Guidelines for the practical use of this approximation are given in Chapter 4.

The vector $\Delta\theta$ inherits the normality from $Y(\theta^*)$ as can easily be seen from (1.13). As a consequence, the *probability density function* (pdf) of $\Delta\theta$ comes close to the normal density:

$$\text{pdf}(\Delta\theta) \approx \sqrt{\frac{\det(J^T(\hat{\theta})J(\hat{\theta}))}{(2\pi\sigma^2)^m}} \exp\left(-\frac{\Delta\theta^T J^T(\hat{\theta})J(\hat{\theta})\Delta\theta}{2\sigma^2}\right) . \quad (1.21)$$

In order to perform a local investigation of (1.10) in the vicinity of the least squares estimate, $\hat{\theta}$, we use a linearisation around $\hat{\theta}$ and the fact that $S(\theta)$ has a minimum at $\theta = \hat{\theta}$, so that:

$$\begin{aligned} S(\hat{\theta} + \Delta\theta) &= Y^T(\hat{\theta} + \Delta\theta)Y(\hat{\theta} + \Delta\theta) \\ &\approx Y^T(\hat{\theta})Y(\hat{\theta}) + \Delta\theta^T (V\Sigma^2V^T) \Delta\theta , \end{aligned} \quad (1.22)$$

where $V = V(\hat{\theta})$ and $\Sigma = \Sigma(\hat{\theta})$ are introduced in (1.15)).

Below we give a brief summary of the statistical background, more details can be found in Chapter 4. A complete treatment of the basic ideas is found in textbooks as [Sch59, DS81, BW88]. According to standard statistics, $S(\hat{\theta})/\sigma^2$ and $\Delta\theta^T(V\Sigma^2V^T)\Delta\theta/\sigma^2$ are independent and have χ^2 -distributions with $N - m$ and m degrees of freedom, respectively. An unbiased estimator of σ^2 is given by

$$s^2 = S(\hat{\theta})/(N - m) . \quad (1.23)$$

The $(1 - \alpha)$ -*confidence region* is the ellipsoidal region

$$\Delta\theta^T (V\Sigma^2V^T) \Delta\theta \leq \frac{m}{N - m} S(\hat{\theta}) \mathcal{F}_\alpha(m, N - m) , \quad (1.24)$$

³The case when this matrix is singular is discussed in Section 4.6.

⁴Notice here the difference between $\delta\theta$, to express a correction during a minimisation process, and $\Delta\theta$, after the minimisation is completed.

where $\mathcal{F}_\alpha(m, N - m)$ is the upper α quantile for *Fisher's F-distribution* with m and $N - m$ degrees of freedom.

The *independent confidence interval* for each estimate is given by:

$$\left[\hat{\theta}_i - \Delta^I \theta_i, \hat{\theta}_i + \Delta^I \theta_i \right], \quad (1.25)$$

with:

$$\Delta^I \theta_i = \sqrt{\frac{m}{N - m} S(\hat{\theta}) \mathcal{F}_\alpha(m, N - m) (V \Sigma^{-2} V^T)_{ii}}.$$

Another quantity often used, but only recommended in combination with independent confidence intervals, is the *dependent confidence interval*:

$$\left[\hat{\theta}_i - \Delta^D \theta_i, \hat{\theta}_i + \Delta^D \theta_i \right], \quad (1.26)$$

with:

$$\Delta^D \theta_i = \sqrt{\frac{m}{N - m} \frac{S(\hat{\theta}) \mathcal{F}_\alpha(m, N - m)}{(V \Sigma^2 V^T)_{ii}}}.$$

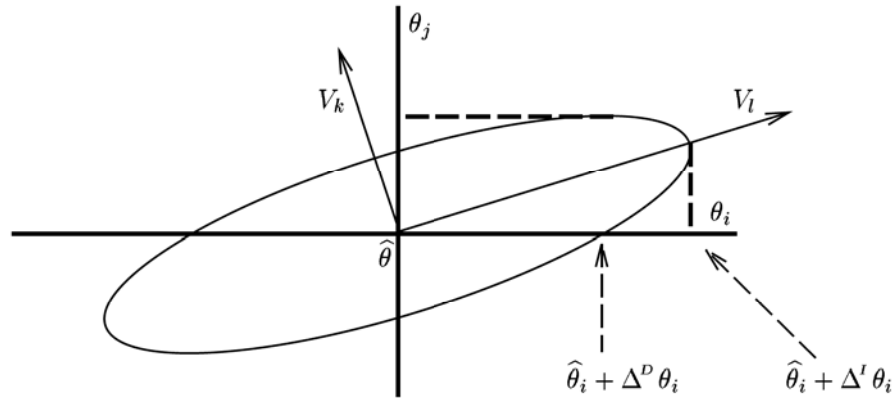


Figure 1.1: Graph of a 2-dimensional intersection of the ellipsoidal region from (1.24), centred at $\hat{\theta}$.

The reader is referred to Figure 1.1 for a graphical interpretation. The principal axes of the *ellipsoidal confidence region* coincide with the column vectors in the matrix V . The distance from the origin to the ellipse along the l -th principle axis (the l -th column of V) is proportional to the reciprocal of the l -th singular value. This means that a small

singular value gives rise to a large confidence region in the direction of the corresponding column vector of V . The independent confidence interval of the i -th parameter (1.25) coincides with the projection of the ellipsoidal region on the i -th parameter axis. The intersection of the ellipse with the i -th parameter axis yields the dependent confidence interval (1.26).

In literature (see for instance [BW88, page 6]) attention is paid to the $(1 - \alpha)$ *marginal confidence region*. Considering only these intervals for the parameters might be misleading, because it does not take into account the correlation between the parameters. This is demonstrated in [DS81, page 95] in the case of an elongated confidence region whose principal axes are not along the axes in the parameter space. In our approach, the ratio of $\Delta^T \theta_i$ and $\Delta^D \theta_i$ indicates this correlation. This ratio is used in Section 6.1.7.

1.7 Parameter constraints

For many practical reasons restrictions may occur with respect to the parameters to be estimated (e.g. reaction constants are always non-negative). Many of the simpler linear restrictions can be taken into account by a reparametrisation, but that is not always possible or even desirable.

Suppose we have K restrictions for the m unknown parameters. The restrictions are, in general, nonlinear and denoted by $R_i(\theta) \leq 0$ for $i = 1, \dots, K$, or

$$R(\theta) \leq 0, \quad (1.27)$$

where $R(\theta)$ is a K -dimensional vector function. The restrictions imply that a subset of the m -dimensional parameter space is excluded. This yields a *constrained minimisation* problem. To solve it, we start the numerical procedure as if we were dealing with the unconstrained case (starting with an initial θ s.t. $R(\theta) \leq 0$) which results in a $\delta\theta$ according to (1.16). Then we check whether after the correction the constraints are still fulfilled: $R(\theta + \delta\theta) \leq 0$. When some of the constraints are violated, there will be a non-empty subset $\mathcal{Z} = \{i_1, \dots, i_k\} \subset \{1, \dots, K\}$, such that $R_j(\theta) > 0$ for $j \in \mathcal{Z}$ and k is the number of violated or active constraints. We end up with a constrained minimisation problem stating: minimise $S(\theta)$ as introduced in (1.4) under the conditions $R_j(\theta) = 0$ for $j \in \mathcal{Z}$.

The first step in solving this constrained minimisation problem is the determination of the above mentioned subset \mathcal{Z} . The second step consists of the computation of the $k \times m$ matrix B defined as:

$$(B)_{jl} = \frac{\partial R_{i_j}}{\partial \theta_l}. \quad (1.28)$$

For notational convenience we introduce a k -dimensional vector $r(\theta)$ which contains all the vector elements $R_{i_j}(\theta)$ for $j \in \{1, \dots, k\}$. If we write down the normal equations with linearised constraints and denote the *Lagrange multipliers* by q , we get:

$$\begin{pmatrix} J^T J & B^T \\ B & \emptyset \end{pmatrix} \begin{pmatrix} \delta\theta \\ q \end{pmatrix} = - \begin{pmatrix} J^T Y(\theta) \\ r(\theta) \end{pmatrix}. \quad (1.29)$$

Making use of the SVD of J , that was already required in the method of Section 1.5, we can easily implement additional parameter constraints in the minimisation procedure. Again we use the Levenberg-Marquardt method to solve the extended nonlinear system. This leads to the correction:

$$\delta\theta = -V (\Sigma^2 + \lambda I_m)^{-1} [\Sigma U^T Y(\theta) + (BV)^T q] , \quad (1.30)$$

where the Lagrange multipliers, q , are given by:

$$q = \left(BV (\Sigma^2 + \lambda I_m)^{-1} (BV)^T \right)^{-1} \left(BV (\Sigma^2 + \lambda I_m)^{-1} \Sigma U^T Y(\theta) - r(\theta) \right) . \quad (1.31)$$

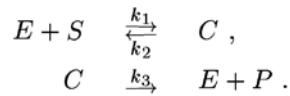
Substitution of (1.31) in (1.30) yields a correction, $\delta\theta$, which satisfies the linearised constraints. It may take some iterations to fulfill all, nonlinear restrictions. Numerical experiments showed that 2 or 3 iterations are usually sufficient. Having found the solution of the constrained minimisation problem, we check the direction of the gradient to be sure that no local minimum is found in the interior; we double-check if all the equality constraints are needed.

In practice, given the constraints (1.27), computer algebra is used to generate the FORTRAN code needed to evaluate the matrix B in (1.28).

1.8 A case study from biochemistry

To illustrate the approach explained in the preceding sections, we consider a simple example in this section. More complex, real-life problems are discussed in Chapter 6.

We consider a simple *enzymatic reaction*, which is a building block for many biochemical processes [Hem72a]. It is given by the chemical equations:



The state variables in the reaction scheme are the concentrations of the enzyme, $[E]$, substrate, $[S]$, and complex, $[C]$. The concentration of the product, $[P]$, is not of interest in this context and therefore not a state variable. The mathematical description of the problem is given by:

$$\begin{aligned} \frac{d[S]}{dt} &= -k_1[E][S] + k_2[C] , \\ \frac{d[C]}{dt} &= k_1[E][S] - k_2[C] - k_3[C] , \\ [E] + [C] &= [E_0] + [C_0] . \end{aligned} \quad (1.32)$$

The initial values are $[S_0] = 1.0$, $[C_0] = 0.0$ and $[E_0] = 1.0$, the vector of unknown, positive parameters is $\theta^T = (k_1, k_2, k_3)$. The data are generated artificially, by adding a

normally distributed, independent measurement error, with zero expectation and fixed variance, to the simulation results of [C]. The resulting complex concentrations are given in Appendix 1.A. As a consequence of the error structure, we take all weights equal in this estimation problem. The initial parameter vector, θ_{ini} , the final estimate, $\hat{\theta}$, the corresponding sum of squared discrepancies (cf. (1.4)) and the confidence limits ($\Delta' \theta$ from (1.25)) are given in Table 1.1. Together with the data, the numerical solution of the DAEs from (1.32) for θ_{ini} and $\hat{\theta}$, is shown in Figure 1.2.

	θ_{ini}	$\hat{\theta}$	$\Delta' \theta$
k_1	6.0	0.683	0.076
k_2	0.8	0.312	0.068
k_3	1.2	0.212	0.005
$S(\theta)$	0.848	0.00051	

Table 1.1: Initial and final parameter values for the case study of Section 1.8 plus $\Delta' \theta$ from (1.25).

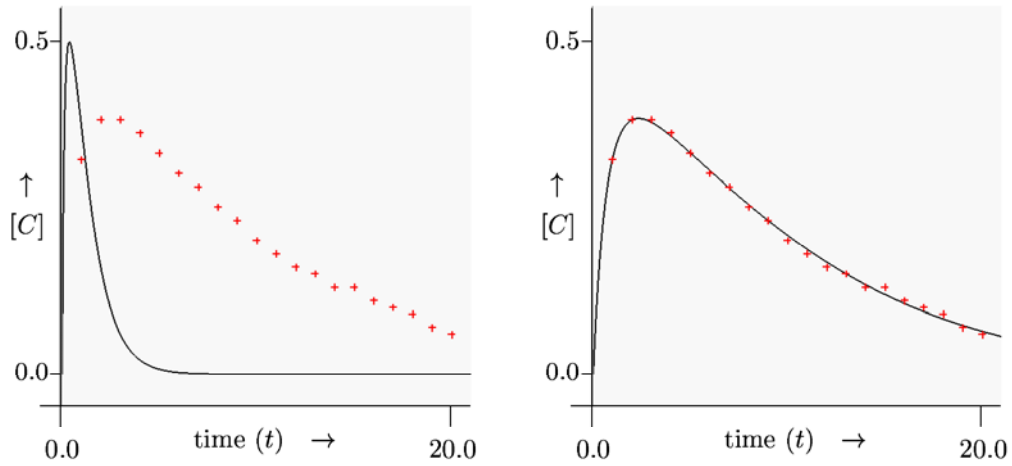


Figure 1.2: The calculated concentrations of the complex ($[C]$) for the initial (left graph) and final parameter vector (right graph) and the measurements (+).

1.9 A case study from population dynamics

Another classical example originates from population dynamics. It describes two species with a *predator-prey* relation. Mathematically the model is described by the *Lotka-Volterra* equations:

$$\frac{dy_1}{dt} = k_1 y_1 - k_2 y_1 y_2, \quad \text{with } y_1(t_0) = y_{1,0}, \quad (1.33)$$

$$\frac{dy_2}{dt} = k_3 y_1 y_2 - k_4 y_2, \quad \text{with } y_2(t_0) = y_{2,0}. \quad (1.34)$$

The rates $k_1 - k_4$ are the parameters to be estimated. A frequently used model adaptation is made by setting the parameters k_2 and k_3 equal to each other. The related estimation problem, with three parameters, is known as *Barnes' problem* and shows up in literature many times as a test example [Hem72b, EW95, HK93, Wik97]. The corresponding measurements can be found in Appendix 1.B.

The initial values of the state variables, $y_{1,0}$ and $y_{2,0}$, are known, but we do not have an indication about their accuracy. We can consider them either as accurate initial conditions or as parameters to be estimated. In the second case we add the given initial values to the measured data. Consequently, we consider four different models to fit the measurements. Statistical tests are performed to discriminate between the candidate models.

When the initial values, $y_{1,0}$ and $y_{2,0}$, are considered as unknown parameters, and k_2 and k_3 are assumed to be two separate, independent parameters, this model fits the data better than the other models which can be derived from (1.33) and (1.34), because the other models can be considered as a special case of this model. With an *F-ratio test* (Appendix 1.C) it can be decided whether one model fits *significantly* better than another. It answers the question: does an increase of the number of parameters lead to a sufficient improvement of the residual sum of squares, $S(\hat{\theta})$?

The degrees of freedom and the corresponding least squares sums for the various models are given in Table 1.2.

Variant	parameters	df. ($N - m$)	$S(\hat{\theta})$
(I)	$y_{1,0}, y_{2,0}, k_1, k_2, k_3, k_4$	22-6	0.05185
(II)	k_1, k_2, k_3, k_4	20-4	0.05592
(III)	$y_{1,0}, y_{2,0}, k_1, k_2, k_4$ ($k_2 = k_3$)	22-5	0.1017
(IV)	k_1, k_2, k_4 ($k_2 = k_3$)	20-3	0.1645

Table 1.2: The parameters, degrees of freedom and the least squares sum for the four proposed variants of the predator-prey model from (1.33) and (1.34).

From this table we can choose $\binom{4}{2} = 6$ pairs of models, 5 of them can be compared by the F-ratio test of the first part of Appendix 1.C. The pair {II,III} is compared by

making use of the super-model I. The 5 pairwise comparisons lead 4 times to a rejection of the null-hypothesis, the F-test on {I,II} did not reject the null-hypothesis. This means that, on the basis of the measurements of Table 1.4, model II is preferred to the other models.

1.10 Concluding remarks

In this chapter we gave an outline of an approach to solve parameter estimation problems in systems of differential algebraic equations. Besides the model equations, which describe the process studied and depend on the unknown parameters, we integrate the corresponding sensitivity equations numerically for an initial guess of the parameter vector. The result forms the input for the minimisation problem, for which we calculate a correction for the parameter vector. For the corrected value the model and sensitivity equations are solved. This iterative process leads to an optimal fit between the model and the data, and the corresponding parameters. After the minimisation the vicinity of the final parameter estimates is investigated in order to derive confidence regions.

The model and variational –or sensitivity– equations are solved numerically by a BDF method, which fully exploits the stiffness character of the variational equations. For the minimisation we use a Levenberg-Marquardt method.

The solution method described has been implemented and can be applied in many sciences where mathematical modelling of time dependent processes is involved. The introductory case studies in this chapter give an impression of the usefulness of the approach. More complicated case studies take up Chapter 6.

Appendix 1.A

The data ($N = 20$) corresponding to the example of Section 1.8 contain simulated values of the complex concentration, with additive, mutually independent errors from a normal distribution. A sequence $\widetilde{[C]}_i$ and the corresponding t_i are given in Table 1.3.

Appendix 1.B

The data, corresponding to Barnes' problem in Section 1.9, for the measured values of the prey and predator fractions, $\widetilde{y}_{1,i}$ and $\widetilde{y}_{2,i}$, respectively, are given in Table 1.4 and taken from [HK93]. The measurements at $t = 0.0$ do not contribute to the number of measurements, N , if the corresponding values are taken as the initial conditions.

Appendix 1.C

We refer to [Rat83] for an introduction to the statistical tests which should be performed and which will help the modeller to decide whether the number of parameters can be

time (t_i)	$\widetilde{[C]}_i$	time (t_i)	$\widetilde{[C]}_i$
1.0	0.32	11.0	0.18
2.0	0.38	12.0	0.16
3.0	0.38	13.0	0.15
4.0	0.36	14.0	0.13
5.0	0.33	15.0	0.13
6.0	0.30	16.0	0.11
7.0	0.28	17.0	0.10
8.0	0.25	18.0	0.09
9.0	0.23	19.0	0.07
10.0	0.20	20.0	0.06

Table 1.3: Measurements of the complex concentration ($[C]$) corresponding to (1.32).

time (t_i)	$\widetilde{y}_{1,i}$	$\widetilde{y}_{2,i}$	time (t_i)	$\widetilde{y}_{1,i}$	$\widetilde{y}_{2,i}$
0.0	1.0	0.30	3.0	0.5	0.30
0.5	1.1	0.35	3.5	0.6	0.25
1.0	1.3	0.40	4.0	0.7	0.25
1.5	1.1	0.50	4.5	0.8	0.30
2.0	0.9	0.50	5.0	1.0	0.35
2.5	0.7	0.40			

Table 1.4: Measurements of prey and predator fractions corresponding to (1.33) and (1.34).

reduced or what model should be chosen. When we have one set of N measurements and two models with approximately the same fit, the model with the fewer parameters is preferred for further investigation. The above notion of ‘approximately’ is made more precise in the remainder of this appendix.

Suppose we have two solutions coming from different models

$$y(t, \theta) \quad \text{and} \quad z(t, \phi). \quad (1.35)$$

We use n_y and m_θ for the dimension of y and θ , respectively. Similarly, the dimensions of z and ϕ are denoted by n_z and m_ϕ . In general, different models describing the same physical process have different numbers of state variables and parameters. The only restriction is that the vectors y and z both contain the state variables for which measurements are available.

Both models have their optimal estimates of the parameters and corresponding residual sums of squares: $\hat{\theta}$, $\hat{\phi}$, $S(\hat{\theta})$ and $S(\hat{\phi})$. From the normality assumption with respect to the measurement errors, and assuming that the optimal estimates of the parameters are close to the true parameter values, we know that the residual sums of squares are approximately χ^2 -distributed:

$$S(\hat{\theta})/\sigma^2 \sim \chi_{N-m_\theta}^2 \quad \text{and} \quad S(\hat{\phi})/\sigma^2 \sim \chi_{N-m_\phi}^2 .$$

It is important to note that the two ratios are dependent, which implies that we cannot perform an F-ratio test straightaway. First, we will consider the case where one model say, $z(t, \phi)$ is a submodel of $y(t, \theta)$. This means that $m_\theta > m_\phi$ and that there exist $m_\theta - m_\phi$ restrictions $h_i(\theta) = 0$, such that $y(t, \theta)$, when it is restricted by $h(\theta) = 0$, has the same input/output behaviour as $z(t, \phi)$ for the observable state variables. Second, we will give an outline of the approach for the case one model is not a submodel of the other one. At the end, we will give an approach which is applicable in both cases, but is more restrictive with respect to N .

In the first case we test the hypothesis: $H_0 : h(\theta) = 0$. Therefore, we consider the ratio $(S(\hat{\phi}) - S(\hat{\theta}))/\sigma^2$, where σ^2 is the variance of the measurement error. This ratio, which is always positive, is independent of $S(\hat{\theta})/\sigma^2$. Now we introduce:

$$X = \frac{(S(\hat{\phi}) - S(\hat{\theta}))/(\sigma^2(m_\theta - m_\phi))}{S(\hat{\theta})/(\sigma^2(N - m_\theta))} \sim \mathcal{F}(m_\theta - m_\phi, N - m_\theta) , \quad (1.36)$$

where $\mathcal{F}(p, q)$ denotes Fisher's F-distribution with p and q degrees of freedom, respectively. From the characteristics of an F-distribution we know:

$$\mathbf{E}(X) = \frac{N - m_\theta}{N - m_\theta - 2} , \quad (\text{for: } N - m_\theta > 2)$$

and

$$P(X \leq \mathcal{F}_\alpha(m_\theta - m_\phi, N - m_\theta)) = 1 - \alpha ,$$

where $\mathcal{F}_\alpha(m_\theta - m_\phi, N - m_\theta)$ is the upper α quantile for Fisher's F-distribution (see e.g. [MGB74]). Notice that the expectation of X does not depend on $m_\theta - m_\phi$. When the two models have about the same performance, X will be close to its expectation. The F-ratio test states that whenever X exceeds $\mathcal{F}_\alpha(m_\theta - m_\phi, N - m_\theta)$, the null-hypothesis, $H_0 : h(\theta) = 0$, should be rejected. If this is the case, then $S(\hat{\phi})$ is significantly larger than $S(\hat{\theta})$, the model which corresponds to $z(t, \phi)$ should be rejected in favour of the model which corresponds to $y(t, \theta)$. When we refer to the F-ratio test in this thesis, we mean this test, unless stated otherwise. Furthermore, we want to stress again that all statements about stochastic behaviour of our statistics hold approximately and are exact only for models which are linear in θ .

In the second case, neither $y(t, \theta)$ is a submodel of $z(t, \phi)$ nor vice versa. Here, we construct a super-model, $u(t, \psi)$, such that $u(t, \psi)$ under the condition $h_\theta(\psi) = 0$ or

$h_\phi(\psi) = 0$ coincides with $y(t, \theta)$ or $z(t, \phi)$, respectively. Because both $y(t, \theta)$ and $z(t, \phi)$ are submodels of $u(t, \psi)$, we return to the first case and compare the models $y(t, \theta)$ and $z(t, \phi)$, by performing the tests with $H_0 : h_\theta(\psi) = 0$ and $H_0 : h_\phi(\psi) = 0$. If one of the two null-hypotheses is rejected, then the submodel corresponding to the non-rejected null-hypothesis is preferred. In all the other cases no conclusion can be drawn.

An approach which is applicable in both cases, if $N \geq 2 \max(m_\theta, m_\phi) + 2$ consists of splitting the data into two disjunct subsets of sizes $N_{(1)}$ and $N_{(2)}$, such that $N_{(1)} + N_{(2)} = N$ and $\min(N_{(1)}, N_{(2)}) \geq \max(m_\theta, m_\phi) + 1$. Then we fit the model $y(t, \theta)$, to the first subset of data, which leads to the estimate $\hat{\theta}_{(1)}$ and the corresponding partial, residual sum $S_{(1)}(\hat{\theta}_{(1)})$. Analogously, we derive $\hat{\theta}_{(2)}$, $\hat{\phi}_{(1)}$, $\hat{\phi}_{(2)}$ and the corresponding partial, residual sums. The null-hypothesis states that the two models perform equally well. Now, we perform two F-ratio tests with:

$$X_{1,2} = \frac{S_{(1)}(\hat{\theta}_{(1)})/(N_{(1)} - m_\theta)}{S_{(2)}(\hat{\phi}_{(2)})/(N_{(2)} - m_\phi)},$$

and

$$X_{2,1} = \frac{S_{(2)}(\hat{\theta}_{(2)})/(N_{(2)} - m_\theta)}{S_{(1)}(\hat{\phi}_{(1)})/(N_{(1)} - m_\phi)}.$$

Consequently, we have:

$$P\left(\frac{1}{\mathcal{F}_{\frac{\alpha}{2}}(N_{(2)} - m_\phi, N_{(1)} - m_\theta)} \leq X_{1,2} \leq \mathcal{F}_{\frac{\alpha}{2}}(N_{(1)} - m_\theta, N_{(2)} - m_\phi)\right) = 1 - \alpha,$$

and

$$P\left(\frac{1}{\mathcal{F}_{\frac{\alpha}{2}}(N_{(1)} - m_\phi, N_{(2)} - m_\theta)} \leq X_{2,1} \leq \mathcal{F}_{\frac{\alpha}{2}}(N_{(2)} - m_\theta, N_{(1)} - m_\phi)\right) = 1 - \alpha.$$

At a confidence level of, at least, $1 - 2\alpha$ we reject the null-hypothesis if one of the F-tests, based on $X_{1,2}$ or $X_{2,1}$, rejects the null-hypothesis in favour of one of the two models and the other test does not contradict this, more precisely:

$$\begin{aligned} &\left(X_{1,2} > \mathcal{F}_{\frac{\alpha}{2}}(N_{(1)} - m_\theta, N_{(2)} - m_\phi) \wedge X_{2,1} \geq \frac{1}{\mathcal{F}_{\frac{\alpha}{2}}(N_{(1)} - m_\phi, N_{(2)} - m_\theta)}\right) \vee \\ &\left(X_{1,2} \geq \frac{1}{\mathcal{F}_{\frac{\alpha}{2}}(N_{(2)} - m_\phi, N_{(1)} - m_\theta)} \wedge X_{2,1} > \mathcal{F}_{\frac{\alpha}{2}}(N_{(2)} - m_\theta, N_{(1)} - m_\phi)\right) \end{aligned} \quad (1.37)$$

or

$$\begin{aligned} &\left(X_{1,2} < \frac{1}{\mathcal{F}_{\frac{\alpha}{2}}(N_{(2)} - m_\phi, N_{(1)} - m_\theta)} \wedge X_{2,1} \leq \mathcal{F}_{\frac{\alpha}{2}}(N_{(2)} - m_\theta, N_{(1)} - m_\phi)\right) \vee \\ &\left(X_{1,2} \leq \mathcal{F}_{\frac{\alpha}{2}}(N_{(1)} - m_\theta, N_{(2)} - m_\phi) \wedge X_{2,1} < \frac{1}{\mathcal{F}_{\frac{\alpha}{2}}(N_{(1)} - m_\phi, N_{(2)} - m_\theta)}\right). \end{aligned} \quad (1.38)$$

If (1.37) is true then $z(t, \phi)$ is chosen in favour of $y(t, \theta)$, the opposite happens if (1.38) is true.

Chapter 2

Parameter Estimation by Total Least Squares

2.1 Introduction

In this chapter we introduce a stable and efficient approach to estimate unknown parameters in nonlinear models where the measurements are affected by noise, not only in the dependent, but also in the independent variables. The technique, where also the error in the independent variable is considered, is known as the *total least squares* (TLS) approach or *errors in variables method* (EVM)¹. A formal, mathematical extension from ordinary (weighted) least squares (OLS) to total least squares (TLS) is found in Section 2.2. Special attention is paid to the consequences of the error structure of the measurements on the parameter estimates in Section 2.3. We restrict ourselves to independent and normally distributed measurement errors whose variances are known or known up to a constant of proportionality. In Section 2.4 we discuss the possibility of adding nonlinear restrictions with respect to the location of the unknown parameters and of adding error margins to the independent variables. A discussion of the case where the variances are unknown or dependences between the measurement errors exist is given in Chapter 3.

Linear TLS problems are discussed in, e.g., [GV83, VPR96], which focus on the numerical linear algebra aspects. Nonlinear problems are discussed in a more theoretical context and with an accent on the statistical context in, for example, [Ful87, Gle90], whereas [ST85, BBS87] focus on the numerical aspects and implementation. This last reference uses the expression orthogonal distance regression. A more complete overview of the topic can be found in the conference proceedings [BF90].

The confidence regions based on the TLS-estimators are not discussed in literature, but will be taken care of in this chapter. With respect to the numerical implementation we will follow a general approach and extend it to the case where parameter constraints and bounds on the measurement errors of the independent variable are given.

In the situation of Figure 2.1 the assumption of an error in the dependent variable, combined with the steep part of the model curve, makes the lack of fit related to the

¹Some texts use the expression orthogonal least squares and abbreviate it by OLS. This might lead to confusion, because the same abbreviation is also used for ordinary least squares (cf. Chapter 1).

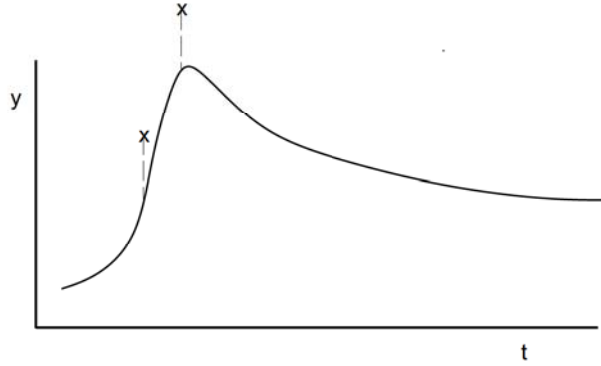


Figure 2.1: A model curve and two measurements: an example where ordinary least squares may be unsatisfactory.

second measurement apparently more significant than the lack of fit related to the first measurement. If we apply the OLS approach here, then only the vertical discrepancies (the dashed lines) are taken into account and both discrepancies will contribute equally to the fitness criterion (1.4) (assuming equal weights). In many problems from various applications the experimentalist will state that the lack of fit in the right measurement is more significant than the lack of fit in the left measurement. These intuitive reasonings lead to a fitness criterion which is more general than OLS.

2.2 Mathematical description of TLS

In most situations one focuses on the case where measurement errors are stochastically independent, come from a normal distribution, and have zero mean and known variance. Further, the errors in the measurements corresponding to the independent variables are assumed to be zero or negligible. The approach needed under these conditions –ordinary least squares– was described in Chapter 1.

In the case when the measurement errors related to the independent variables are significant we need the more general TLS approach. Using OLS in such cases is called the naive approach in [Gle90] and leads to biased, inconsistent estimators. For some applications, e.g., curve fitting, OLS may not even lead to an estimate, whereas TLS does.

As we want to consider a possible measurement error with respect to the independent variable, t , we have to adapt our notation for a measurement as given in (1.2). Now a *measurement* is denoted by the triple:

$$(c_i, \tilde{t}_i, \tilde{y}_i), \quad i = 1, \dots, N, \quad (2.1)$$

where, the measured time, \tilde{t}_i , replaces the actual time of the measurement, t_i , the symbols c_i and \tilde{y}_i have the same meaning as in (1.2). The fitness criterion of (1.4), is not appropriate any more, because the error in \tilde{t}_i may be significant and –more importantly– t_i is not known. The naive approach would be to replace t_i by \tilde{t}_i and use a least squares criterion.

For the measurement errors in time, ξ_i ($i = 1, \dots, N$), which are assumed to be $\mathcal{N}(0, \zeta_i^2)$ distributed and stochastically independent, we write:

$$\tilde{t}_i = t_i + \xi_i , \quad (2.2)$$

where the actual or true times of the measurements, t_i , are not known. The discrepancies related to the independent variable are denoted by τ_i , such that for the true model $\tau^* = (\tau_1^*, \dots, \tau_N^*)^T = (-\xi_1, \dots, -\xi_N)^T$. An estimator of the error in time is denoted by $\hat{\tau}$. As a consequence, the discrepancy between the measured value and the theoretical value of a dependent variable now depends on θ and τ :

$$d_i(\theta, \tau_i) = y_{c_i}(\tilde{t}_i + \tau_i, \theta) - \tilde{y}_i . \quad (2.3)$$

After adding weights, the expression we want to minimise reads:

$$S(\theta, \tau) = \sum_{i=1}^N w_i^2 \{ d_i^2(\theta, \tau_i) + v_i^2 \tau_i^2 \} . \quad (2.4)$$

Here, w_i is a weighting factor for the i -th measurement and v_i represents a weighting factor, with dimension $[y/t]$, which indicates the relative importance of τ_i compared to d_i . At this stage we assume the weights, w_i and v_i , to be known a priori.

For convenience we introduce the following notation:

$$\begin{aligned} \nu &= \begin{pmatrix} \theta \\ \tau \end{pmatrix} , \\ g_i(\nu) &= \begin{cases} w_i(y_{c_i}(\tilde{t}_i + \tau_i, \theta) - \tilde{y}_i) = w_i d_i(\theta, \tau_i) , & i = 1, \dots, N , \\ v_{i-N} w_{i-N} \tau_{i-N} , & i = N+1, \dots, 2N , \end{cases} \\ g(\nu) &: \mathbb{R}^{(m+N)} \rightarrow \mathbb{R}^{2N} , \\ S(\nu) &= g^T(\nu) g(\nu) , \\ Z &= \frac{dg}{d\nu} . \end{aligned}$$

This notation is used to describe a numerical procedure to minimise $S(\nu)$. The computation of the discrepancies and sensitivities is performed by the same means as in Chapter 1, with the only difference that the evaluations take place at $\tilde{t}_i + \tau_i$. The initial estimate, ν_{ini} , equals $(\theta_{ini}, 0, \dots, 0)^T$, where θ_{ini} is the initial guess for the parameters as introduced in Chapter 1, for τ_{ini} we take its expected zero vector. At this stage, we can focus on the computation of an optimal solution by numerical means. In principle, this

can be done by the *Gauss-Newton* method. In each iteration we compute a correction for ν , denoted by $\delta\nu$, from the *normal equations*

$$Z^T Z \delta\nu = -Z^T g(\nu) . \quad (2.5)$$

In order to compute $\delta\nu$ efficiently and to investigate the differences with the minimisation from Section 1.5, we analyse the $2N \times (m + N)$ -matrix A by partitioning this matrix as:

$$Z = \begin{pmatrix} J & C \\ 0 & D \end{pmatrix} , \quad (2.6)$$

with:

$$(J)_{ij} = \frac{\partial g_i}{\partial \theta_j} = w_i \frac{\partial y_{c_i}(\tilde{t}_i + \tau_i, \theta)}{\partial \theta_j} , \quad (2.7)$$

$$(i = 1, \dots, N, \quad j = 1, \dots, m) ,$$

$$(C)_{ij} = \frac{\partial g_i}{\partial \tau_j} = w_i \frac{\partial y_{c_i}(\tilde{t}_i + \tau_i, \theta)}{\partial \tau_j} , \quad (2.8)$$

$$(i = 1, \dots, N, \quad j = 1, \dots, N) ,$$

$$(D)_{ij} = \frac{\partial g_{i+N}}{\partial \tau_j} = \delta_{ij} v_i w_i , \quad (2.9)$$

$$(i = 1, \dots, N, \quad j = 1, \dots, N) ,$$

where δ_{ij} is the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j , \\ 0 & \text{otherwise.} \end{cases}$$

If a BDF method is used to solve the model equations (1.1) numerically, the entries of C are retrieved easily. A closer look at C and D shows that these matrices are both diagonal. Substitution of (2.6) in (2.5) and partitioning $g(\nu)$ into its two components, $g^\theta(\nu)$ and $g^\tau(\nu)$, both of length N , yields:

$$\begin{pmatrix} J^T J & J^T C \\ C J & D^2 + C^2 \end{pmatrix} \begin{pmatrix} \delta\theta \\ \delta\tau \end{pmatrix} = - \begin{pmatrix} J^T & 0 \\ C & D \end{pmatrix} \begin{pmatrix} g^\theta(\nu) \\ g^\tau(\nu) \end{pmatrix} . \quad (2.10)$$

Because of the diagonal structure of the matrices C and D , it is obviously easiest to start with the lower half of (2.10) and compute the correction:

$$\delta\tau = - (D^2 + C^2)^{-1} [C (J\delta\theta + g^\theta(\nu)) + Dg^\tau(\nu)] , \quad (2.11)$$

which, after substitution in the upper half of (2.10), leads to the expression for $\delta\theta$:

$$\begin{aligned} J^T (I - C(D^2 + C^2)^{-1} C) J \delta\theta &= -J^T [I - C(D^2 + C^2)^{-1} C] g^\theta(\nu) + \\ &\quad J^T C (D^2 + C^2)^{-1} D g^\tau(\nu) . \end{aligned} \quad (2.12)$$

In order to get a more convenient notation we introduce the diagonal $N \times N$ -matrix T such that:

$$T^2 = (I - C(D^2 + C^2)^{-1}C) , \quad (2.13)$$

followed by introducing:

$$\tilde{J} = TJ , \quad (2.14)$$

$$h = Tg^\theta(\nu) - T^{-1}C(D^2 + C^2)^{-1}Dg^\tau(\nu) . \quad (2.15)$$

With this notation we simply express the normal equations for $\delta\theta$ (cf. (1.13)) by:

$$\tilde{J}^T \tilde{J} \delta\theta = -\tilde{J}^T h . \quad (2.16)$$

Notice that TLS reduces to OLS if C vanishes. Equation (2.16) can be solved by the Levenberg-Marquardt method as described in Chapter 1, which only needs a slight adaptation. After computing $\delta\theta$ from (2.16), by making an SVD of \tilde{J} , the result is substituted into (2.11) to obtain $\delta\tau$. Thus, the Levenberg-Marquardt method is not applied to (2.5), but to the smaller problem (2.16), which has the same size as the problem in the OLS case. The matrix multiplication to obtain \tilde{J} and the substitution which has to be made to calculate $\delta\tau$ are marginal computations compared to computing $\delta\theta$ from (2.16). This means that TLS takes about the same amount of computational time as OLS and is therefore solved in an efficient way. Furthermore, the numerical solution is similar to the solution of the OLS approach and therefore the stability and the convergence are the same as for OLS.

Notice that in the derivation of the above formulae we assume the weights, w_i and v_i , to be known a priori.

2.3 Statistical background

In this section we assume the measurement errors in the independent and dependent variables, τ^* and $d(\theta^*, \tau^*)$ respectively, to be stochastically independent, normally distributed and scaled by their weights in such a way that the covariance matrix is given by:

$$\mathbf{E} \left(\begin{pmatrix} g^\theta(\nu^*) \\ g^\tau(\nu^*) \end{pmatrix} \begin{pmatrix} g^\theta(\nu^*) \\ g^\tau(\nu^*) \end{pmatrix}^T \right) = \sigma^2 I_{2N} , \quad (2.17)$$

where ν^* contains the true parameter values. This assumption means that the standard deviation of every measurement error is proportional to the reciprocal of its weight, i.e. $\sigma_i = \sigma/w_i$ and $\zeta_i = \sigma/(v_i w_i)$. This is a matter of scaling and we need these conditions to ensure that the total least squares estimate coincides with the maximum likelihood estimate (MLE) as discussed in more detail in Chapter 3.

Minimisation of $S(\nu)$ leads to a final estimate of the unknown parameters ν , denoted by $\hat{\nu}$. Combining the normal equations from (2.10) and the covariance matrix of the measurement errors (2.17) leads to the approximate covariance matrix of $\Delta\nu = \nu^* - \hat{\nu}$:

$$\begin{aligned} \mathbf{E}(\Delta\nu\Delta\nu^T) &= \mathbf{E} \left(\begin{pmatrix} \Delta\theta \\ \Delta\tau \end{pmatrix} \begin{pmatrix} \Delta\theta \\ \Delta\tau \end{pmatrix}^T \right) = \sigma^2 \begin{pmatrix} J^T J & J^T C \\ C J & C^2 + D^2 \end{pmatrix}^{-1} = \\ &\sigma^2 \begin{pmatrix} (\tilde{J}^T \tilde{J})^{-1} & -(\tilde{J}^T \tilde{J})^{-1} J^T C (C^2 + D^2)^{-1} \\ -(C^2 + D^2)^{-1} C J (\tilde{J}^T \tilde{J})^{-1} & (C^2 + D^2)^{-1} [I_N + C J (\tilde{J}^T \tilde{J})^{-1} J^T C (C^2 + D^2)^{-1}] \end{pmatrix}, \end{aligned} \quad (2.18)$$

where the last expression only contains known inverses. As in (1.22) we perform a local investigation of the sum of squares in the vicinity of the final estimate, $\hat{\nu}$, by using a linear approximation for $g(\hat{\nu} + \Delta\nu)$:

$$\begin{aligned} S(\hat{\nu} + \Delta\nu) &= g^T(\hat{\nu} + \Delta\nu)g(\hat{\nu} + \Delta\nu) \\ &\approx g^T(\hat{\nu})g(\hat{\nu}) + \Delta\nu^T Z^T Z \Delta\nu, \end{aligned} \quad (2.19)$$

where the matrix Z is given in (2.6) and evaluated at $\hat{\nu}$.

At this point we apply standard statistics as in Section 1.6, but have to be careful about counting the degrees of freedom. The criterion to be minimised, $S(\nu)$, is the sum of $2N$ squared discrepancies. At the minimum $dS(\nu)/d\nu = 0$ holds, which leads to $N + m$ restrictions. As a result, $S(\hat{\nu})/\sigma^2$ and $\Delta\nu^T Z^T Z \Delta\nu/\sigma^2$ have χ^2 -distributions with $N - m$ and $N + m$ degrees of freedom, respectively. The confidence region at level α is the ellipsoidal region

$$\Delta\nu^T Z^T Z \Delta\nu \leq \frac{N + m}{N - m} S(\hat{\nu}) \mathcal{F}_\alpha(N + m, N - m), \quad (2.20)$$

where $\mathcal{F}_\alpha(N + m, N - m)$ denotes the upper α quantile for Fisher's F-distribution with $N + m$ and $N - m$ degrees of freedom. From this last result, which is an extension of the standard linear regression theory, individual confidence regions for each estimate can be calculated as in (1.25) and (1.26), respectively.

An approximately unbiased estimator of σ^2 is given by

$$s^2 = S(\hat{\nu})/(N - m). \quad (2.21)$$

2.4 Total least squares with parameter constraints

In this section we study the case where, in addition to the minimisation criterion, a set of constraints with respect to ν is given. The approach to handle this situation is an extension of Section 1.7. Using the notation of Section 2.2 with respect to g, ν and $S(\nu)$, we state the *constrained minimisation* problem as:

$$\min_{\nu} g^T(\nu)g(\nu), \quad \text{under the condition: } R(\nu) \leq 0, \quad (2.22)$$

with: $R : \mathbb{R}^{(m+N)} \rightarrow \mathbb{R}^K$ denoting K nonlinear constraints. We assume $R(\nu)$ to be differentiable with respect to ν . We start the numerical procedure in the case of constrained minimisation as if we were dealing with the unconstrained case (starting with an initial estimate of ν satisfying the constraints $R(\nu) \leq 0$), which results in a $\delta\nu$. Then we check if, after a correction of ν , the constraints are still satisfied:

$$R(\nu + \delta\nu) \leq 0 .$$

If this is the case, we do not have to worry about the restrictions and continue with the next iteration as if it were an unconstrained minimisation problem. If some of the K constraints are violated, there will be a subset $\mathcal{Z} = \{i_1, \dots, i_k\} \subset \{1, \dots, K\}$, such that $R_j > 0$ ($j \in \mathcal{Z}$), where k denotes the number of active constraints.

After determining the subset \mathcal{Z} , we compute the $k \times m$ matrix B_1 and the $k \times N$ matrix B_2 , defined as:

$$(B_1)_{jl} = \frac{\partial R_{i_j}}{\partial \theta_l} \quad \text{and} \quad (B_2)_{jl} = \frac{\partial R_{i_j}}{\partial \tau_l} . \quad (2.23)$$

In the software these matrices are derived automatically via a *computer algebra* package (we used MAPLE). For notational convenience we introduce a k -dimensional vector $r(\nu)$ which contains all vector elements R_j for $j \in \mathcal{Z}$. If we write down the normal equations with linearised constraints and denote the *Lagrange multipliers* by q , we obtain:

$$\begin{pmatrix} J^T J & J^T C & B_1^T \\ C J & D^2 + C^2 & B_2^T \\ B_1 & B_2 & 0 \end{pmatrix} \begin{pmatrix} \delta\theta \\ \delta\tau \\ q \end{pmatrix} = - \begin{pmatrix} J^T g^\theta(\nu) \\ C g^\theta(\nu) + D g^\tau(\nu) \\ r(\nu) \end{pmatrix} . \quad (2.24)$$

In the remainder of this section we show how (2.24) can be solved by making use of the special structure of the matrices of these normal equations and of preparatory computations with respect to J . We start by writing $\delta\tau$ explicitly:

$$\delta\tau = -(D^2 + C^2)^{-1} (C g^\theta(\nu) + D g^\tau(\nu) + C J \delta\theta + B_2^T q) , \quad (2.25)$$

and substitute this in the first row of equation (2.24):

$$\begin{aligned} -J^T C (D^2 + C^2)^{-1} (C g^\theta(\nu) + D g^\tau(\nu) + C J \delta\theta + B_2^T q) + \\ J^T J \delta\theta + B_1^T q = -J^T g^\theta(\nu) , \end{aligned} \quad (2.26)$$

which can be rewritten as:

$$\begin{aligned} J^T (I_N - C (D^2 + C^2)^{-1} C) J \delta\theta &= -J^T (I_N - C (D^2 + C^2)^{-1} C) g^\theta(\nu) + \\ J^T C (D^2 + C^2)^{-1} D g^\tau(\nu) &+ (J^T C (D^2 + C^2)^{-1} B_2^T - B_1^T) q . \end{aligned} \quad (2.27)$$

Using the matrices T and \tilde{J} , and h as from (2.13)-(2.15), we find:

$$\delta\theta = - \left(\tilde{J}^T \tilde{J} \right)^{-1} \left(\tilde{J}^T h + \{ B_1^T - J^T C (D^2 + C^2)^{-1} B_2^T \} q \right) , \quad (2.28)$$

where the SVD of $\tilde{J}^T \tilde{J}$ is available, because we started as if we were dealing with the unconstrained case and therefore had to solve (2.16) already.

Finally, pre-multiplying the equations (2.25) and (2.28) by B_2 and B_1 respectively, adding the two results and eliminating $\delta\theta$ via (2.28), we can use the last row of equation (2.24) to obtain:

$$\begin{aligned} & \left[\{B_1 - B_2(D^2 + C^2)^{-1}CJ\} (\tilde{J}^T \tilde{J})^{-1} \right. \\ & \left. \{J^T C(D^2 + C^2)^{-1}B_2^T - B_1^T\} - B_2(D^2 + C^2)^{-1}B_2^T \right] q = \\ & \{B_1 - B_2(D^2 + C^2)^{-1}CJ\} (\tilde{J}^T \tilde{J})^{-1} \tilde{J}^T h + \\ & B_2(D^2 + C^2)^{-1}[Cg^\theta(\nu) + Dg^\tau(\nu)] - r(\nu) . \end{aligned} \quad (2.29)$$

The last equation is solved to obtain q , its size is governed by the number of violated constraints, k . For most applications this number is small, which means that the Lagrange multipliers, q , can be solved easily and fast from system (2.29), e.g., by a QR-decomposition. After the computation of q , the correction $\delta\theta$ can be computed by (2.28) and $\delta\tau$ from (2.25). As in the OLS case, at the end a set of equations with the size of the number of violated constraints has to be solved. For the TLS case we have marginal extra work for extra multiplications and additions, the time consuming parts, solving q from (2.29) and performing the SVD of an $N \times m$ -matrix stay the same for the OLS and the TLS approach.

2.5 Conclusions

In this chapter we presented an approach for parameter estimation in nonlinear models, where not only the measurement errors in the dependent, but also in the independent variables have to be taken into account. This approach is known as the total least squares (TLS) method in contrast to the ordinary least squares (OLS) approach, where the measurement errors in the independent variables are neglected. We showed how to deal with nonlinear restrictions with respect to the unknown parameters and error bounds of the independent variables. Special attention was paid to confidence regions of the final estimates.

The TLS approach is more general than the OLS approach and it reduces to OLS in a natural way, if the weighted errors in the independent variable are negligible. The increase in the computational effort for the TLS approach is marginal compared to the OLS approach.

Chapter 3

Maximum Likelihood Estimators

3.1 Introduction

In this chapter we give a more detailed description of the statistical background for parameter estimation in nonlinear models, also known as *nonlinear regression*. The fitness criteria used in nonlinear regression depend on the assumptions and knowledge about the measurement errors. From the probability density function of the measurement error the maximum likelihood estimates of the parameters can be derived. For the case with independent and normally distributed measurement errors in the dependent variables, we discuss the link between least squares and maximum likelihood criteria in the Sections 3.2 and 3.3. An outline of the actual optimisation of these criteria by numerical means, when the variances of the measurement error are unknown, is considered in Section 3.4. A theoretical outline concerning dependent measurement errors with an unknown covariance matrix is given in Section 3.5, the consequences for actual computation are highlighted in Section 3.6.

Maximum likelihood methods for the case when the measurement errors are normally distributed and also present in the independent variable are discussed in Section 3.7. When the measurement errors come from a Laplace –or double exponential– distribution, the sum of absolute discrepancies should be minimised. Section 3.8 gives the necessary background and an elegant way to deal with the practical implementation. Concluding remarks can be found in Section 3.9.

Throughout this chapter, we assume that an accurate approximation of the solution of the model and its variational equations, $y(t, \theta)$ and $\partial y(t, \theta)/\partial \theta$ (cf. (1.6)) is available and we do not bother about the precise formulation of the model.

3.2 Least squares criterion

The most straightforward way to measure the fitness between the model and the measurements is the sum of squared discrepancies:

$$S(\theta) = \sum_{i=1}^N (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2 = \sum_{i=1}^N d_i^2(\theta) = d^T(\theta)d(\theta) , \quad (3.1)$$

where $d(\theta) = (d_1(\theta), d_2(\theta), \dots, d_N(\theta))^T$.

Assuming that all measurement errors, ε_i , are mutually independent and come from a normal or Gaussian distribution, with zero mean and variance σ^2 , i.e. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, the vector of measurement errors reads: $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)^T$, with covariance matrix

$$V = \mathbf{E}(\varepsilon \varepsilon^T) = \sigma^2 I_N . \quad (3.2)$$

The discrepancies, $d(\theta) \in \mathbb{R}^N$, depend on the parameter vector. When the true parameter vector, θ^* , is substituted, the discrepancies coincide with the measurement errors:

$$\tilde{y}_i = y_{c_i}(t_i, \theta^*) + \varepsilon_i \quad \text{or} \quad d_i(\theta^*) = -\varepsilon_i .$$

By *residuals* we mean the discrepancies evaluated for the estimated parameter vector, $d(\hat{\theta})$. The *probability density function* for the assumed structure of the measurement errors, is given by:

$$\begin{aligned} p(\tilde{y}_1, \dots, \tilde{y}_N | \theta) &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left(-\frac{\sum_{i=1}^N (y_{c_i}(t_i, \theta) - \tilde{y}_i)^2}{2\sigma^2} \right) \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left(-\frac{\sum_{i=1}^N d_i^2(\theta)}{2\sigma^2} \right) \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left(-\frac{1}{2} d^T(\theta) V^{-1} d(\theta) \right) . \end{aligned} \quad (3.3)$$

We want to determine θ in such a way that the probability density is maximal, i.e. the most likely θ , for a given data set. From the probability density function we can define the *likelihood function* as:

$$\mathcal{L}(\theta) \equiv p(\theta | \tilde{y}_1, \dots, \tilde{y}_N) . \quad (3.4)$$

For convenience and convention we take the logarithm of the likelihood function (*LLF*):

$$\ln \mathcal{L}(\theta) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} d^T(\theta) V^{-1} d(\theta) . \quad (3.5)$$

The likelihood function (and its logarithm) reaches its maximum, if $S(\theta)$ in (3.1) is minimal because of (3.2). This means that the maximum likelihood (ML) estimate of θ coincides with its least squares (LS) estimate. As a consequence, the last sentence can be expressed as

$$\hat{\theta}_{\text{ML}} \stackrel{\text{def}}{=} \{\theta | \mathcal{L}(\theta) \text{ is maximal}\} = \hat{\theta}_{\text{LS}} \stackrel{\text{def}}{=} \{\theta | S(\theta) \text{ is minimal}\} ,$$

where $\hat{}$ indicates an estimate.

3.3 Weighted least squares criterion

3.3.1 A priori known weights

In the case some a priori knowledge about the accuracy of the measurements is available and this accuracy is not constant over the components of the state vector or even differs for two measurements of the same component, an adaptation of the criterion function has to be made. The expression for $S(\theta)$ in (3.1) is changed by adding positive weights, w_i ($i = 1, \dots, N$), which leads to a sum of weighted squared discrepancies (1.4). The weights are taken in such a way that more accurate measurements correspond to bigger weights.

If we assume again that the errors are independent and come from a Gaussian distribution with non-constant variance, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, the corresponding logarithm of the likelihood function reads:

$$\ln \mathcal{L}(\theta) = -\frac{N}{2} \ln(2\pi) - \sum_{i=1}^N \ln(\sigma_i) - \frac{1}{2} \sum_{i=1}^N \left(\frac{d_i}{\sigma_i} \right)^2 . \quad (3.6)$$

After comparing (1.4) and (3.6), we see that their estimates coincide if and only if, the weights are proportional to the reciprocal of the standard deviations:

$$w_i = \frac{\sigma}{\sigma_i} , \quad (3.7)$$

which connects weighted least squares and maximum likelihood estimates for the case of non-constant variances.

If the measurement errors are dependent and the covariance matrix is known, $\varepsilon \sim \mathcal{N}(0, V)$, with V a symmetric, non-diagonal, positive definite $N \times N$ -matrix, we use a more general LLF instead of (3.5):

$$\ln \mathcal{L}(\theta) = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(V)) - \frac{1}{2} d^T(\theta) V^{-1} d(\theta) , \quad (3.8)$$

whose maximum coincides with the minimum of:

$$S(\theta) = d^T(\theta) V^{-1} d(\theta) . \quad (3.9)$$

Due to the properties of V , the matrix V^{-1} can be decomposed by Cholesky factorisation, such that $V^{-1} = L^T L$, where L is a lower triangular matrix. With this matrix L , the problem can be transformed into a least squares problem, almost similar to the one in Chapter 1. In literature, the minimisation of (3.9) is known as the generalised least squares (GLS) problem.

3.3.2 Unknown weights

In most practical situations the standard deviations of the measurement errors, σ_i , are unknown. Furthermore, it is impossible to estimate all these standard deviations, in addition to the unknown parameters, θ . We exclude the possibility of a systematic error,

so that the expectation of the measurement error is assumed to vanish. The best that can be done is to assume that the measurement errors come from the same distribution if they correspond to the same component of the state vector.¹ This means that besides the unknown parameters we estimate as many standard deviations as different components, c_i in (1.2), have been measured.

We introduce q as the number of measured components, $q \leq n$, and r as the number of samples. A set of measurements for different components, c_i , taken at the same time and under the same experimental conditions builds a *sample*. We define the $r \times q$ matrix $D(\theta)$ containing the discrepancies, $d_l(\theta)$, in such a way that each column is associated with one measured component and each row corresponds to one sample. We adapt the notation of (3.1) correspondingly and use a double subscript for the entries of the matrix $D(\theta)$ instead of the single index we use for $d_l(\theta)$. The entry $D_{ij}(\theta)$ corresponds to the j -th measurement of the i -th sample. Notice that some entries of the matrix $D(\theta)$ may be empty, because it may happen that $N < qr$. At these empty entries we put a zero. Thus, there is a one-to-one correspondence between $d_l(\theta)$ ($l = 1, \dots, N$), and the N non-zero entries in the matrix $D(\theta)$.

With the matrix $D(\theta)$ we introduce the $q \times q$ matrix $M(\theta)$, given by:

$$M(\theta) = D^T(\theta)D(\theta) . \quad (3.10)$$

In literature ([Bar74, page 64]), $M(\theta)$ is known as the *moment matrix*. Although both $D(\theta)$ and $M(\theta)$ depend on the unknown parameter vector θ , we will not always express this dependence in the notation.

Until Section 3.5 we are dealing with stochastically independent measurement errors. This, together with the assumption that the deviations, σ_i , are the same for each measured component, turns V into a diagonal, $q \times q$ -matrix, with $V_{ii} = \sigma_i^2$, $i = 1, \dots, q$.

The introduction of M and V results in a shorthand notation for the weighted sum of squares. Instead of (1.4), we get:

$$S(\theta) = \text{Tr}(V^{-1}M) , \quad (3.11)$$

where Tr denotes the trace of a matrix. Starting with the special case where the same components are measured in each sample and hence $N = qr$, we will conclude with the more general case at the end of this section. For this special case the probability density function reads:

$$\begin{aligned} p(\tilde{y}_1, \dots, \tilde{y}_N | \theta) &= \left(\prod_{j=1}^q (2\pi)^{-\frac{r}{2}} \right) \left(\prod_{j=1}^q \prod_{i=1}^r \left(\frac{1}{V_{jj}} \right)^{\frac{1}{2}} \right) \exp(-\frac{1}{2} \text{Tr}(V^{-1}M)) \\ &= (2\pi)^{-\frac{qr}{2}} \left(\prod_{j=1}^q \left(\frac{1}{V_{jj}} \right)^{\frac{r}{2}} \right) \exp(-\frac{1}{2} \text{Tr}(V^{-1}M)) \end{aligned}$$

¹This is the approach for absolute measurement errors, in the case of relative measurement errors the situation is identical after scaling the measurement errors.

$$= (2\pi)^{-\frac{qr}{2}} (\det(V^{-1}))^{\frac{r}{2}} \exp(-\frac{1}{2}\text{Tr}(V^{-1}M)) . \quad (3.12)$$

The corresponding log likelihood function equals:

$$\begin{aligned} \ln \mathcal{L}(\theta) &= -qr \ln(\sqrt{2\pi}) + \sum_{j=1}^q \frac{r}{2} \ln \left(\frac{1}{V_{jj}} \right) - \frac{1}{2} \text{Tr}(V^{-1}M) \\ &= -\frac{r}{2} \left(q \ln(2\pi) + \sum_{j=1}^q \ln(V_{jj}) \right) - \frac{1}{2} \sum_{j=1}^q \frac{1}{V_{jj}} \sum_{i=1}^r D_{ij}^2 . \end{aligned} \quad (3.13)$$

Differentiation with respect to the unknown variances of the LLF from the last equation gives:

$$\frac{\partial(\ln \mathcal{L}(\theta))}{\partial V_{jj}} = -\frac{r}{2V_{jj}} + \frac{1}{2V_{jj}^2} \sum_{i=1}^r D_{ij}^2 ,$$

which vanishes iff:

$$V_{jj} = \frac{1}{r} \sum_{i=1}^r D_{ij}^2 = \frac{1}{r} M_{jj} .$$

Inspection shows that the resulting stationary point corresponds to a maximum. The last equation yields an estimator of the variances, which is consistent, but biased. Consistency is easy to show; when $N \rightarrow \infty$, then also $r \rightarrow \infty$, because q is finite and bounded by n , and finally, by the law of large numbers:

$$P \left(\left| \hat{V}_{jj} - V_{jj}^* \right| > \epsilon \right) \rightarrow 0 , \quad \forall \epsilon > 0 ,$$

where V_{jj}^* is the true variance. As an approximately unbiased and consistent estimator we take, according to [Bar74, page 195]:

$$\hat{V}_{jj} = \frac{1}{r(1 - m/N)} \sum_{i=1}^r D_{ij}^2 = \frac{1}{r(1 - m/N)} M_{jj} . \quad (3.14)$$

this estimator is perfectly unbiased if the expectation of the matrix M is proportional to V^* . The adaptation in the denominator expresses that the degrees of freedom are spread over the separate entries of the estimator. With respect to the last equation a special remark should be made. To estimate the diagonal matrix V^* we use the diagonal entries of M . For the estimator, $\hat{\theta}$, the residual, $D(\hat{\theta})_{ij}$, is expected to come from a normal distribution with zero expectation, and variance σ_j^2 . Therefore, the off-diagonal entries $M_{ij}(\hat{\theta})$ ($i \neq j$) are expected to have a zero expectation and a variance $r\sigma_i^2\sigma_j^2$, if the measurement errors are independent. These characteristics can be used to test whether the combination of the model chosen and the assumption of the independent measurement errors is feasible.

The result (3.14) holds only if all measured components are the same over the samples, $N = qr$. For the more general case we introduce the variables r_j to denote the number of measurements in the j -th column of D ($\sum_{j=1}^q r_j = N$). Then, the LLF reads

$$\ln \mathcal{L}(\theta) = -\frac{N}{2} \ln(2\pi) - \sum_{j=1}^q \frac{r_j}{2} \ln(V_{jj}) - \frac{1}{2} \sum_{j=1}^q \frac{1}{V_{jj}} \sum_{i=1}^r D_{ij}^2, \quad (3.15)$$

its derivative with respect to V_{jj} vanishes if:

$$V_{jj} = \frac{1}{r_j} \sum_{i=1}^r D_{ij}^2. \quad (3.16)$$

The corresponding approximately, unbiased, consistent estimate of the variances is – analogous to (3.14) – given by

$$\hat{V}_{jj} = \frac{1}{r_j(1 - m/N)} \sum_{i=1}^r D_{ij}^2. \quad (3.17)$$

Notice that the summation runs over r entries, because of the zeros substituted in the matrix D .

Substitution of (3.17) in (3.15) gives:

$$\begin{aligned} \ln \mathcal{L}(\theta) = & -\frac{N}{2} \ln(2\pi) - \sum_{j=1}^q \frac{r_j}{2} \ln \left(\frac{1}{r_j(1 - \frac{m}{N})} \right) - \sum_{j=1}^q \frac{r_j}{2} \ln \left(\sum_{i=1}^r D_{ij}^2 \right) - \\ & \frac{1}{2} \sum_{j=1}^q \frac{r_j(1 - \frac{m}{N})}{\sum_{i=1}^r D_{ij}^2} \sum_{i=1}^r D_{ij}^2. \end{aligned} \quad (3.18)$$

Only the third term in the right-hand side of (3.18) depends on θ , which means that we can restrict ourselves to minimising:

$$\tilde{\mathcal{L}}(\theta) = \prod_{j=1}^q \left(\sum_{i=1}^r D_{ij}^2 \right)^{\frac{r_j}{2}}. \quad (3.19)$$

From (3.19) we see that we have to minimise the geometric mean of the estimated deviations of the measurement error, where we omit the factor $1/(r_j(1 - m/N))$. Another interpretation is to consider an N -dimensional box in the data space. This box is centred at the expected model responses and has edges parallel to the coordinate axes. The length of the edge parallel to the l -th coordinate axis is proportional to $\sqrt{\sum_{i=1}^r D_{il}^2}$, where the j -th column of D corresponds with the measured component c_l . Minimising the volume of this box is expressed by (3.19). In the next section we describe how to perform this minimisation.

3.4 Numerical computation (independent case)

To compute the maximum likelihood estimates for θ in the case of independent measurement errors and unknown weights, the expression (3.19) should be minimised. This might be done by any general purpose minimisation routine. Newton's method would be a straightforward procedure if accurate initial estimates would be available, but problems are expected due to the strong nonlinear behaviour of this criterion. Another disadvantage of direct minimisation of (3.19) is the fact that its first and second derivatives, which might be required by the minimisation routine, lead to more complex expressions than in the case of, for example, ordinary least squares. In order to obtain the estimates we introduce an alternative iteration procedure which is a slightly modified least squares approach. Therefore, it is easy to adapt an existing approach as described in Chapter 1, where no adaptations for the derivatives have to be made. The alternative approach proves to be applicable, efficient and stable in all practical cases.

The proposed approach to find a minimum of (3.19) is an iterative procedure. The process starts with the solution of the model and variational equations as described in Chapter 1 for a given initial estimate of the parameter vector, θ_{ini} , and possibly additional constraints on the parameter vector as also introduced in that chapter. During the iterative process this computation is repeated with different weights, which depend on θ in the way as given below.

In order to explain the successive computations we use the iteration index k . At the k -th step of the minimisation procedure the parameter vector is given by $\theta^{(k)}$, so that $\theta^{(0)} = \theta_{ini}$ and $D_{ij}(\theta^{(k)})$ denotes the corresponding discrepancies, which are known after computing the model responses. Estimates of the variances at this stage are given by:

$$\hat{\sigma}_j^2(\theta^{(k)}) = \frac{1}{r_j} \sum_{i=1}^r D_{ij}^2(\theta^{(k)}) , \quad (3.20)$$

which is the biased estimate from (3.16). Notice that the biased and the approximately unbiased estimates for V^* only differ by a proportionality factor, which has no influence on the final estimate of θ . If the weights in (1.4) are chosen as in (3.7) and we take (3.20) as the estimate for σ_j^2 , the corresponding weighted sum of squares, cf. (3.11), reads:

$$S(\theta^{(k)}) = \sum_{j=1}^q \frac{1}{\hat{\sigma}_j^2(\theta^{(k)})} \sum_{i=1}^r D_{ij}^2(\theta^{(k)}) = N . \quad (3.21)$$

Now we continue the procedure as if in (3.21) only the discrepancies, and not the variances, depend on θ . I.e., we compute a new $\theta^{(k+1)}$ for an adapted set of weights, $w_j = 1/\hat{\sigma}_j(\theta^{(k)})$. A correction for $\theta^{(k)}$, denoted by $\delta\theta^{(k)}$, is accepted, if it leads to an improvement of the sum of squares with the delayed or frozen weights:

$$S(\theta^{(k+1)}) := \sum_{j=1}^q \frac{1}{\hat{\sigma}_j^2(\theta^{(k)})} \sum_{i=1}^r D_{ij}^2(\theta^{(k+1)}) < N , \quad (3.22)$$

where

$$\theta^{(k+1)} = \theta^{(k)} + \delta\theta^{(k)} .$$

After a successful correction, the weights are updated and the next iteration is performed. For the iterative minimisation we use the Levenberg-Marquardt algorithm as in Section 1.5.

Thus by introducing a weighted least squares problem of type (1.4), where the weights lag behind over the iteration steps, we manage to create a process for minimising (3.19). In the remainder of this section we show that the iteration from (3.22) leads to the minimisation of (3.19) at a superlinear convergence rate.

Theorem 3.4.1 The value $\hat{\theta}$, corresponding to a stationary point of $S(\theta)$ of the iteration process (3.22) minimises the value $\tilde{\mathcal{L}}(\theta)$ in (3.19). Moreover, if the residual is sufficiently small and if the derivatives $\partial^2 y_{c_i} / \partial \theta^2(t_i, \theta^{(k)})$ and $\partial^3 y_{c_i} / \partial \theta^3(t_i, \theta^{(k)})$ exist for all k steps of the iteration, then the rate of convergence of (3.22) is superlinear.

Proof: First we consider the iterative process as described in (3.22). For the correction we get an expression which is common for such processes as:

$$\delta\theta^{(k)} = \theta^{(k+1)} - \theta^{(k)} = -W^{-1}(\theta^{(k)})Z(\theta^{(k)}) , \quad (3.23)$$

where $W(\theta^{(k)})$ is an $m \times m$ -matrix and, depending on the local minimisation method, equal to or approximating the Hessian, $(\partial^2 S / \partial \theta^2)(\theta^{(k)})$ and $Z(\theta^{(k)})$ an approximate gradient vector $(\partial S / \partial \theta)(\theta^{(k)})$. Because we ‘freeze’ the variances, in our algorithm the first derivative of (3.22) with respect to θ equals:

$$Z(\theta) = \sum_{j=1}^q \frac{1}{\hat{\sigma}_j^2} \sum_{i=1}^r 2D_{ij} \frac{\partial D_{ij}}{\partial \theta} . \quad (3.24)$$

The gradient of (3.19) reads:

$$\frac{\partial \tilde{\mathcal{L}}(\theta)}{\partial \theta} = \tilde{\mathcal{L}}(\theta) \left(\sum_{j=1}^q \frac{r_j}{\sum_{i=1}^r D_{ij}^2} \sum_{i=1}^r D_{ij} \frac{\partial D_{ij}}{\partial \theta} \right) . \quad (3.25)$$

Upon convergence of (3.23), the correction $\delta\theta^{(k)}$ vanishes, and therefore, also the difference in the weights vanishes. This implies that the expression $r_j / \sum_{i=1}^r D_{ij}^2$ in the right-hand side of (3.25) equals $1/\hat{\sigma}_j^2$ and, thus, the zeros of (3.25) coincide with those of (3.24).

In order to investigate the convergence rate of the iterative procedure, we introduce:

$$F(\theta) = W^{-1}(\theta)Z(\theta) ,$$

so that the converged parameter vector, $\hat{\theta}$, is characterised by $Z(\hat{\theta}) = 0$. We denote the error in the k -th iteration step by:

$$e^{(k)} = \hat{\theta} - \theta^{(k)} .$$

For the errors the following recursive relation holds:

$$e^{(k+1)} = e^{(k)} - \left(F'(\hat{\theta}) - F'(\theta^{(k)}) \right) e^{(k)} .$$

Expanding this relation for small $\|e^{(k)}\|$, we find:

$$\left\| e^{(k+1)} - \left(I - F'(\theta^{(k)}) \right) e^{(k)} \right\| = \mathcal{O} \left(\|e^{(k)}\|^2 \right) , \quad (3.26)$$

where:

$$F'(\theta^{(k)}) = \frac{dF(\theta)}{d\theta}(\theta^{(k)}) = W^{-1}(\theta^{(k)})W(\theta^{(k)}) + \frac{dW^{-1}(\theta)}{d\theta}(\theta^{(k)})Z(\theta^{(k)}) .$$

If $\theta^{(k)}$ in limit goes to $\hat{\theta}$, the gradient $Z(\theta^{(k)})$ vanishes and therefore, $F'(\theta^{(k)})$ in limit goes to the identity matrix. This means that the process has a superlinear convergence rate. \square

Remark 3.4.1 Because $\tilde{\mathcal{L}}(\theta) > 0$, the derivatives (3.25) and (3.24) have identical signs and therefore the functions $S(\theta)$ and $\tilde{\mathcal{L}}(\theta)$ have the same type of stationary points.

Remark 3.4.2 If the matrix W does not contain second order derivatives of $y_{c_i}(t, \theta)$ with respect to θ , as in the case of Gauss-Newton type methods, then the restriction on the third derivative of $y_{c_i}(t, \theta)$ with respect to θ becomes redundant.

3.5 Dependent measurement errors

In the case of dependent measurement errors with unknown dependences, we consider a full, symmetric positive definite, $q \times q$ covariance matrix V . Whereas, in the case of independent measurement errors only the q entries on the diagonal have to be estimated, for the dependent case $q(q+1)/2$ entries are unknown. These unknown quantities come in addition to the m unknown parameters of the vector θ .

Again, we start with the special case that all measured components are the same over the samples, $qr = N$, the corresponding LLF can be rewritten from (3.13) as follows:

$$\ln \mathcal{L}(\theta) = -N \ln \left(\sqrt{2\pi} \right) - \frac{r}{2} \ln(\det(V)) - \frac{1}{2} \text{Tr}(V^{-1}M) . \quad (3.27)$$

Notice that depending on the statistical assumptions, V is either a $q \times q$ - or an $N \times N$ -matrix, the corresponding LLFs are given by (3.13) and (3.8), respectively. In order

to differentiate (3.27) with respect to the entries of the matrix V , we summarise the following results (see [Bar74, pages 294–296]):

$$\frac{\partial \det(A)}{\partial A_{ij}} = (A^{-1})_{ji} \det(A) , \quad (3.28)$$

$$\frac{\partial \text{Tr}(BA^T C)}{\partial A_{ij}} = (CB)_{ij} , \quad (3.29)$$

$$\frac{\partial A_{kl}^{-1}}{\partial A_{ij}} = -A_{ki}^{-1} A_{jl}^{-1} . \quad (3.30)$$

Now the second term in (3.27) can be differentiated with respect to V by using the result (3.28). The derivative of the last term in (3.27) with respect to V can be obtained by combining (3.29) and (3.30). The result of differentiating (3.27) with respect to the covariance matrix reads:

$$\frac{\partial (\ln \mathcal{L}(\theta))}{\partial V} = -\frac{r}{2} V^{-1} + \frac{1}{2} V^{-1} M V^{-1} . \quad (3.31)$$

This expression vanishes if:

$$V = \frac{1}{r} M . \quad (3.32)$$

The last expression gives a consistent, but biased estimator of the covariance matrix. Analogous to (3.14), a less biased estimator is given by:

$$\hat{V} = \frac{1}{r(1 - m/N)} M . \quad (3.33)$$

If we substitute this estimator of the covariance matrix in the LLF (3.27) we obtain:

$$\begin{aligned} \ln \mathcal{L}(\theta) &= -\frac{N}{2} \ln(2\pi) - \frac{r}{2} \ln \left(\left(\frac{1}{r - m/q} \right)^q \det(M) \right) - \frac{1}{2} \text{Tr}((r - m/q) I_q) \\ &= \frac{N}{2} \ln \left(\frac{N - m}{2q\pi} \right) - \frac{r}{2} \ln(\det(M)) + \frac{1}{2}(m - N) . \end{aligned}$$

Maximising this expression with respect to θ is equivalent to minimising:

$$\tilde{\mathcal{L}}(\theta) = \det(M) . \quad (3.34)$$

Due to the relation between the moment matrix, M , and the estimator of the covariance matrix, \hat{V} (cf. (3.33)), we see that minimising (3.34) leads to minimising the volume of an N -dimensional box in the data space. In the case of independent measurement errors, the edges of this box are parallel to the coordinate axes in the data space. In the case of dependent measurement errors the box will have a different orientation. If the covariance matrix is not known, we minimise the volume of this box. The

minimisation is done, not only by adapting the lengths of the edges, but we also allow the box to rotate in the data space.

Analogous to the case of independent measurement errors, we consider the case where qr exceeds N . To this end we introduce the matrices V_i , $i = 1, \dots, r$. The matrix V_i , corresponding to the i -th sample, can be derived from the covariance matrix, V , by omission of the j -th row and the j -th column for each j which has not been measured in the i -th sample. The resulting likelihood function equals

$$\mathcal{L}(\theta) = (2\pi)^{(N/2)} \left(\prod_{i=1}^r (\det(V_i))^{-\frac{1}{2}} \right) \exp\left(-\frac{1}{2} \text{Tr}(V^{-1}M)\right) \quad (3.35)$$

and its logarithm

$$\ln \mathcal{L}(\theta) = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^r \ln(\det(V_i)) - \frac{1}{2} \text{Tr}(V^{-1}M) . \quad (3.36)$$

Minimisation of one of these two expressions is not essentially more difficult than the minimisation of (3.34), but when the method is implemented in an algorithm the non-equal sample sizes should be taken into account.

3.6 Numerical computation (dependent case)

In the case of dependent measurement errors, instead of (3.19), we have to minimise (3.34), which is the determinant of a full, symmetric positive definite matrix. Its dimension equals the number of measured components of the state vector, $y(t, \theta)$. Essentially this minimisation is realised by a method analogous to the technique introduced in Section 3.4. We introduce an approach with a delayed covariance matrix and show that this leads to the minimisation of (3.34).

The optimal parameters are computed by an iterative procedure. Starting at $k = 0$ and an initial guess for the parameters, $\theta^{(0)} = \theta_{ini}$, we solve the model equations, calculate the discrepancies, $d_i(\theta^{(0)})$, and form the matrix M as described in (3.10).

At the k -th step of the iterative minimisation, the estimate of the covariance matrix is given by

$$\hat{V}(\theta^{(k)}) = \frac{1}{r(1 - m/N)} M(\theta^{(k)}) . \quad (3.37)$$

For the final estimates of the parameters it makes no difference if we use a biased or an approximately unbiased estimate for the covariance matrix, because the minimisation is not affected by multiplying $M(\theta^{(k)})$ with a scalar. During an iteration step the estimate of the covariance matrix, $\hat{V}(\theta^{(k)})$, is frozen. We compute a corrected parameter vector, $\theta^{(k+1)} = \theta^{(k)} + \delta\theta^{(k)}$, such that the adapted LLF:

$$\ln \tilde{\mathcal{L}}(\theta^{(k+1)}) = -\frac{N}{2} \ln(2\pi) - \frac{r}{2} \ln \left(\det \left(\hat{V}(\theta^{(k)}) \right) \right) - \frac{1}{2} \text{Tr} \left(\hat{V}^{-1}(\theta^{(k)}) M(\theta^{(k+1)}) \right)$$

is maximal, which is the same as minimising:

$$\begin{aligned} \mathcal{S}(\theta^{(k+1)}) &= \text{Tr} \left(\hat{V}^{-1}(\theta^{(k)}) M(\theta^{(k+1)}) \right) \\ &= \sum_{i=1}^q \sum_{j=1}^q \left(\hat{V}^{-1}(\theta^{(k)}) \right)_{ij} \sum_{l=1}^r D_{li}(\theta^{(k+1)}) D_{lj}(\theta^{(k+1)}) . \end{aligned} \quad (3.38)$$

Instead of minimising the determinant of a matrix as in (3.34), we have transformed the problem into a least squares problem as in (3.9). The additional computation consists of a Cholesky factorisation of $\hat{V}(\theta^{(k)})$ and calculation of its inverse. This computation is not prohibitive, because the matrix $\hat{V}(\theta^{(k)})$ is small for practical cases (we did not encounter real-life problems with $q \geq 10$). Further, the matrix is expected to have the larger entries to be found on the diagonal due to the expected small dependences between the measurement errors.

In the remainder of this section we will prove that $\mathcal{S}(\theta^{(k+1)})$ of (3.38) has the same stationary points as $\tilde{\mathcal{L}}(\theta)$ of (3.34).

Theorem 3.6.1 The iterative procedure, consisting of a sequence of quadratic minimisation problems for $\mathcal{S}(\theta^{(k+1)})$, as described in (3.38) and the minimisation of $\ln \tilde{\mathcal{L}}(\theta)$ from (3.34) reach their stationary points for identical values of θ . The rate of convergence of the iterative procedure is superlinear, under the same conditions as in Theorem 3.4.1.

Proof: We consider the gradients of $\ln \mathcal{S}(\theta)$ and $\ln \tilde{\mathcal{L}}(\theta)$. Differentiation of (3.34) and the use of (3.28) yields:

$$\begin{aligned} \frac{\partial \ln \tilde{\mathcal{L}}(\theta)}{\partial \theta} &= \frac{\partial \ln(\det(M))}{\partial \theta} = \\ \sum_{i=1}^q \sum_{j=1}^q \frac{\partial \ln(\det(M))}{\partial M_{ij}} \frac{\partial M_{ij}}{\partial \theta} &= \sum_{i=1}^q \sum_{j=1}^q (M^{-1})_{ij} \frac{\partial M_{ij}}{\partial \theta} . \end{aligned} \quad (3.39)$$

The same procedure for (3.38) by making use of (3.29), where it is kept in mind that the matrix V is kept fixed in every step of the iteration and therefore does not depend on θ , leads to:

$$\begin{aligned} \frac{\partial \ln \mathcal{S}(\theta)}{\partial \theta} &= \frac{\partial \text{Tr}(V^{-1}M)}{\partial \theta} = \frac{\partial \text{Tr}(V^{-1}M)}{\partial M} \frac{\partial M}{\partial \theta} = \\ \sum_{i=1}^q \sum_{j=1}^q (V^{-1})_{ij} \frac{\partial M_{ij}}{\partial \theta} &= 2 \sum_{i=1}^q \sum_{j=1}^q (V^{-1})_{ij} \sum_{l=1}^r \frac{\partial D_{li}}{\partial \theta} D_{lj} . \end{aligned} \quad (3.40)$$

Upon convergence of the iterative procedure, the correction and therefore the lag of $\hat{V}(\theta^{(k)})$ vanish. As a consequence the matrices M from (3.39) and V in (3.40), are the same up to a scalar factor. This means that the zeros of the derivatives coincide, which completes the first part of the proof.

The proof of the superlinear convergence rate is completely analogous to the proof of Theorem 3.4.1 and is therefore omitted. \square

3.7 MLE and total least squares

In the previous sections we showed under which conditions an ordinary (weighted) least squares approach (OLS) yields maximum likelihood estimates, and how to deal with an unknown covariance matrix. A more general approach for the case a measurement error is also present in the independent variable (TLS) is described in this section.

The notation here is adopted from Section 2.2 and will be extended in Section 3.7.2 in order to deal with a more general situation. We start in Section 3.7.1 with independent measurement errors and a priori known weights. Unknown weights are considered in Section 3.7.2. In Section 3.7.3, we assume independence of the measurement errors and finally in Section 3.7.4 we consider dependent measurement errors.

3.7.1 A priori known weights

First we investigate the probability density corresponding to TLS and derive a condition under which a TLS estimate coincides with the MLE. In Chapter 2 we assumed that the measurement errors are scaled by their weights such that they all may be considered as coming from a $\mathcal{N}(0, \sigma^2)$ distribution. Here we drop this assumption and start with:

$$\left. \begin{aligned} \varepsilon_i &= -d_i(\theta^*, \tau_i^*) \sim \mathcal{N}(0, \sigma_i^2) \\ \xi_i &= -\tau_i^* \sim \mathcal{N}(0, \zeta_i^2) \\ \mathbf{E}(\varepsilon_i, \xi_i) &= 0 \end{aligned} \right\} \quad (i = 1, \dots, N) .$$

$$\mathbf{E}(\varepsilon_i, \varepsilon_j) = \mathbf{E}(\xi_i, \xi_j) = 0 \quad (i, j = 1, \dots, N \text{ and } i \neq j)$$

Taking this error structure into account, the corresponding probability density reads:

$$\begin{aligned} & p(\tilde{y}_1, \dots, \tilde{y}_N, \tilde{t}_1, \dots, \tilde{t}_N | \theta, \tau) \\ &= \left(\frac{1}{2\pi} \right)^{\frac{N}{2}} \prod_{i=1}^N \frac{1}{\sigma_i} \exp \left(-\frac{1}{2} \sum_{i=1}^N \frac{(y_{ci}(\tilde{t}_i + \tau_i, \theta) - \tilde{y}_i)^2}{\sigma_i^2} \right) \\ & \times \left(\frac{1}{2\pi} \right)^{\frac{N}{2}} \prod_{i=1}^N \frac{1}{\zeta_i} \exp \left(-\frac{1}{2} \sum_{i=1}^N \frac{\tau_i^2}{\zeta_i^2} \right) \\ &= \left(\frac{1}{2\pi} \right)^N \prod_{i=1}^N \frac{1}{\sigma_i \zeta_i} \exp \left(-\frac{1}{2} \sum_{i=1}^N \left\{ \frac{d_i^2(\theta, \tau)}{\sigma_i^2} + \frac{\tau_i^2}{\zeta_i^2} \right\} \right) . \end{aligned} \quad (3.41)$$

Analogous to Section 3.3.1, we consider the log likelihood function, LLF:

$$\ln(\mathcal{L}(\theta, \tau)) = -N \ln(2\pi) - \sum_{i=1}^N \ln(\sigma_i \zeta_i) - \frac{1}{2} \sum_{i=1}^N \left\{ \frac{d_i^2(\theta, \tau)}{\sigma_i^2} + \frac{\tau_i^2}{\zeta_i^2} \right\} . \quad (3.42)$$

Inspection of (2.4) and (3.42) shows that their estimates for θ and τ coincide iff:

$$w_i = \frac{\sigma}{\sigma_i} \quad \text{and} \quad v_i w_i = \frac{\sigma}{\zeta_i} , \quad (3.43)$$

which is in accordance with the result of (3.7). The relation of (3.43) shows under which conditions the sum of total least squares and the maximum likelihood function lead to the same estimates.

3.7.2 Unknown weights (TLS)

If the weights are not a priori known, we have to adapt our notation with respect to the discrepancies. As in Section 3.3.2. we construct the $r \times q$ -matrix D . In the same way as D contains the discrepancies for measurements related to the dependent variables, we introduce an $r \times q$ -matrix Ψ , which contains the discrepancies for the independent variable, τ_i . The corresponding *moment matrix* (cf. (3.10) for the OLS case) becomes the $2q \times 2q$ matrix:

$$M = [D|\Psi]^T[D|\Psi] . \quad (3.44)$$

For the same reasons as explained in Section 3.3.2 we assume that variances and covariances do not depend on the time of the measurement, but depend only on the measured component. The $2q \times 2q$ covariance matrix, whose diagonal elements represent the variances, is denoted by V . The non-diagonal elements of V represent the covariances of the measurement errors.

After this introduction of the matrices M and V , the maximum likelihood function can be written as:

$$\mathcal{L}(\theta, \tau) = (2\pi)^{-N} \det(V^{-1})^{\frac{r}{2}} \exp(-\frac{1}{2} \text{Tr}(V^{-1}M)) . \quad (3.45)$$

The maximum likelihood estimates (MLEs) are those values of θ and τ which maximise this expression.

3.7.3 Independent measurement errors

For unknown weights and independent measurement errors the covariance matrix, V , is diagonal and its elements are given by: $\sigma_1^2, \dots, \sigma_q^2, \zeta_1^2, \dots, \zeta_q^2$. The likelihood function in this case is given by:

$$\mathcal{L}(\theta, \tau) = (2\pi)^{-N} \prod_{j=1}^q \left(\frac{1}{\sigma_j \zeta_j} \right)^r \exp \left(-\frac{1}{2} \sum_{j=1}^q \left\{ \frac{1}{\sigma_j^2} \sum_{i=1}^r D_{ij}^2 + \frac{1}{\zeta_j^2} \sum_{i=1}^r \Psi_{ij}^2 \right\} \right), \quad (3.46)$$

and the corresponding LLF reads:

$$\begin{aligned} \ln \mathcal{L}(\theta, \tau) &= -N \ln(2\pi) + \sum_{j=1}^q r \ln \left(\frac{1}{\sigma_j \zeta_j} \right) \\ &\quad - \frac{1}{2} \sum_{j=1}^q \left(\frac{1}{\sigma_j^2} \sum_{i=1}^r D_{ij}^2 + \frac{1}{\zeta_j^2} \sum_{i=1}^r \Psi_{ij}^2 \right) . \end{aligned} \quad (3.47)$$

Computing the maximum of this expression with respect to the variances, σ_j^2 and ζ_j^2 , we get the most likely variances:

$$\sigma_j^2 = V_{jj} = \frac{\sum_{i=1}^r D_{ij}^2}{r}, \quad j \in \{1, \dots, q\}, \quad (3.48)$$

and

$$\zeta_j^2 = V_{q+j, q+j} = \frac{\sum_{i=1}^r \Psi_{ij}^2}{r}, \quad j \in \{1, \dots, q\}. \quad (3.49)$$

Substitution of (3.48) and (3.49) in equation (3.47) leads, after some rewriting, to:

$$\ln \tilde{\mathcal{L}}(\theta, \tau) = \sum_{j=1}^q \ln \left(\sum_{i=1}^r D_{ij}^2 \sum_{i=1}^r \Psi_{ij}^2 \right), \quad (3.50)$$

which is the final criterion function we have to minimise. For the actual minimisation we follow the same strategy as described in Section 3.4.

3.7.4 Dependent measurement errors

Now we drop the assumption with respect to the independence of the measurement errors, although we still assume a normal distribution. Consequently, we now have a full and unknown covariance matrix. Therefore, besides the m unknown parameters from the vector θ , and N measurement errors in the independent variable, denoted by the vector τ , a matrix with $q(2q+1)$ unknown entries has to be estimated.

The general likelihood function (for full matrices V) was given by (3.45). For convenience we take the corresponding LLF to maximise:

$$\ln \mathcal{L}(\theta, \tau) = -N \ln(2\pi) - \frac{r}{2} \ln(\det(V)) - \frac{1}{2} \text{Tr}(V^{-1}M). \quad (3.51)$$

Annihilating the derivative with respect to the elements of the matrix V , we obtain the most likely covariance matrix. Differentiation yields, the same formula as (3.31) –but now with an extended meaning–:

$$\frac{\partial \ln \mathcal{L}}{\partial V} = -\frac{r}{2} V^{-1} + \frac{1}{2} V^{-1} M V^{-1}, \quad (3.52)$$

which vanishes for $V = \frac{1}{r} M$. In order to obtain the final estimates, $\hat{\theta}$ and $\hat{\tau}$, we substitute $V = \frac{1}{r} M$ in the MLE of (3.45). Consequently, we have to minimise:

$$\tilde{\mathcal{L}} = \det(M).$$

As in Section 3.6 the minimisation can be achieved by an iterative process, where the covariance matrix lags behind.

3.8 L_1 -optimisation and Laplace distribution

At the end of this chapter we consider the case where the measurement errors come from a double exponential or Laplace distribution. For convenience we only consider weights that are a priori known.

The probability density function corresponding to measurement errors from a Laplace distribution is given by:

$$p(\tilde{y}_1, \dots, \tilde{y}_N | \theta) = \prod_{i=1}^N \frac{1}{2\sigma_i} \exp\left(-\frac{|d_i(\theta)|}{\sigma_i}\right), \quad (3.53)$$

which leads to the LLF:

$$\ln \mathcal{L}(\theta) = -\sum_{i=1}^N \ln(2\sigma_i) - \sum_{i=1}^N \frac{|d_i(\theta)|}{\sigma_i}. \quad (3.54)$$

Thus, the corresponding function to minimise is:

$$S(\theta) = \sum_{i=1}^N w_i |d_i(\theta)|, \quad (3.55)$$

where the weights are positive and the discrepancies are as defined in (3.1). The estimates of (3.55) and (3.54) coincide if and only if $\sigma_i = \sigma/w_i$, where σ is a proportionality factor. The same relation between the weights and the deviations was also derived in (3.7) in the case of normal measurement errors. It shows that measurement errors from a Laplace distribution lead to an L_1 -optimisation problem.

A method which uses the fitness criterion (3.55) is known to be less sensitive to outliers. This property is called robustness in statistical terminology. The main disadvantage of (3.55) is the discontinuity of the derivative. As a consequence, these methods generally require more sophisticated numerical techniques.

An alternative fitness criterion, which is also more robust than weighted least squares is the Huber M-estimator [Hub81, HW94]. This estimator is defined as the minimum of:

$$T(\theta) = \sum_{i=1}^N \psi(w_i d_i / v), \quad (3.56)$$

where v is a scaling factor and

$$\psi(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq 1, \\ |x| - \frac{1}{2} & |x| > 1. \end{cases}$$

This alternative formulation is differentiable, but second order derivatives do not exist for $x = \pm 1$. This means that, e.g., Newton's method cannot be used and the actual minimisation contains many checks on the bounds of $\psi(w_i d_i / v)$. Therefore, this approach

via (3.56) is less straightforward than a least squares criterion, although numerically easier to tackle than (3.55).

We want to combine the best of both methods: a method which is not too sensitive to outliers and can be implemented easily. To our opinion a simple and reliable remedy can be used here. We use a similar technique as introduced earlier in this chapter, when we used delayed weights. For the computation of L_1 estimates we introduce an iterative process. First, we rewrite (3.55) as:

$$S(\theta) = \sum_{i=1}^N \frac{w_i^2 d_i^2(\theta)}{w_i |d_i(\theta)|}. \quad (3.57)$$

Subsequently, we start an iterative procedure and freeze the denominator, which leads to:

$$S(\theta^{(k)} + \delta\theta^{(k)}) = \sum_{i=1}^N \frac{w_i^2 d_i^2(\theta^{(k)} + \delta\theta^{(k)})}{w_i |d_i(\theta^{(k)})|}. \quad (3.58)$$

This iterative process converges at a superlinear rate. The derivation of this convergence rate is similar as in Theorem 3.4.1 and hence is omitted. The minimisation problem of (3.57) can be solved with a standard least squares minimisation routine, such as Gauss-Newton or Levenberg-Marquardt. The denominator of (3.58) needs some special care to avoid numerical instabilities. We choose to add a threshold value to the denominator in order to prevent division by zero. Consequently, weighted discrepancies which are smaller than this threshold, inliers, get a smaller weight. This is not a reason for concern because the contribution of these inliers to the sum of absolute discrepancies is marginal, with or without this threshold.

3.9 Conclusions

In this chapter we presented maximum likelihood estimates (MLEs) for measurement errors from a Gaussian and a Laplace distribution. We explained the links with least squares, total least squares and L_1 -optimisation, under different assumptions about the knowledge and the structure of covariance matrix.

Numerical methods were introduced to calculate these estimates. They appear to be stable and are attractive because of their good convergence properties and relatively simple implementation once a reliable algorithm for weighted least squares is available.

In the case the error structure is a priori known in detail, it is a valuable exercise to neglect this information on the error structure and to investigate if, e.g., the a posteriori calculated (estimated) covariance matrix is in agreement with the one a priori known. Discrepancy between the expected and estimated structure of the measurement errors is a good starting point for model adaptations or a review on the statistical assumptions with respect to the measurement errors.

Chapter 4

Nonlinear Regression, Bias and Curvature

4.1 Overview of the chapter

In this chapter we give an overview of some aspects of the theory of nonlinear regression, which have practical relevance when physical models are calibrated. Not only the computation of the parameter estimates, but also the statistical properties of the corresponding estimator depend –besides the error structure of the measurements– heavily on the nonlinearity of the regression problem. In this chapter we discuss the consequences of nonlinearity when a least squares estimation criterion is used.

We start with a short overview of the theory for linear regression in Section 4.2. From this overview we will look into the differences between the linear and nonlinear case. Sections 4.3 and 4.4 contain a number of approaches to quantify the nonlinearity of a regression problem. Bias measures for the parameters contain information about the separate parameters, but do not indicate whether this nonlinearity can be reduced by a reparametrisation. The curvature measures of Section 4.4 make a distinction between intrinsic and parameter dependent nonlinearity.

Nonlinearity measures can be derived by either analytic means or by computationally intensive means. We will compare their performances and discuss the advantages and disadvantages of both approaches. The choice of a certain approach depends also on the underlying model and the time it takes to calculate an accurate model response.

We conclude this chapter with a collection of related problems such as sampling techniques on and graphical representations of levelsets in Section 4.5, and the consequences of parameter constraints on level sets and over-parametrisation in Section 4.6.

In this chapter we assume that not only accurate approximations of $y(t, \theta)$ and $\partial y(t, \theta)/\partial \theta$ are available, but also sufficiently accurate approximations of $\partial^2 y(t, \theta)/\partial \theta^2$. The latter will be used to derive analytic measures for the extent of nonlinearity.

4.2 Linear Regression

A thorough overview of the theory of linear regression can be found in standard texts as, e.g., [DS81, Rao73, Sch59, Seb77]. We just give a brief overview with the aim to introduce the necessary notation:

- $t \in \mathbb{R}$ is the regressor, explanatory or independent variable,
- $y \in \mathbb{R}^n$ is the vector of response or dependent variables,
- $\theta \in \mathbb{R}^m$ is the vector of unknown parameters to be estimated.

The fact that we deal with one independent variable only is not a restriction; t can be replaced by an $x \in \mathbb{R}^l$ without further consequences. In the case of linear regression, the regression function is linear in the unknown parameters, θ , written as:

$$y(t, \theta) = X(t) \theta, \quad (4.1)$$

where X is an $n \times m$ -matrix independent of θ , but depending –possibly nonlinearly– on t .

A set of measurements is denoted by triples as in (1.2). For the true parameter vector, θ^* , we have

$$\tilde{y}_i = y_{c_i}(t_i, \theta^*) = X_{c_i}(t_i) \theta^* + \varepsilon_i, \quad i = 1, \dots, N, \quad (4.2)$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and independent of ε_j (for $i \neq j$)¹, and $X_{c_i}(t_i)$ is the c_i -th row of X , evaluated at t_i . Notice that $N \geq m$ is a necessary condition in order to be able to determine an estimator for all m parameters. The weighted least squares estimate, $\hat{\theta}$, minimises the weighted sum of squared discrepancies. The corresponding criterion reads:

$$S(\theta) = \sum_{i=1}^N w_i^2 (X_{c_i}(t_i) \theta - \tilde{y}_i)^2 = Y^T(\theta) Y(\theta), \quad (4.3)$$

where $Y(\theta)$ is an N -dimensional vector containing the weighted discrepancies. The derivative of (4.3) with respect to θ equals:

$$\frac{\partial S}{\partial \theta} = 2 \frac{\partial Y^T(\theta)}{\partial \theta} Y(\theta) = 2 J^T Y(\theta), \quad (4.4)$$

with the elements of the Jacobian, J , given by:

$$(J)_{ij} = w_i X_{c_{ij}}(t_i), \quad i = 1, \dots, N, \quad j = 1, \dots, m.$$

The minimum of (4.3) is attained for $\hat{\theta}$, the solution of the normal equations:

$$(J^T J) \hat{\theta} = J^T (w_1 \tilde{y}_1, w_2 \tilde{y}_2, \dots, w_N \tilde{y}_N)^T. \quad (4.5)$$

¹In Section 3.3.1 we showed that the more general case, $\varepsilon \sim \mathcal{N}(0, V)$, where V is a symmetric, positive definite matrix, can be reduced to this generic case.

Clearly, $\text{Rank}(J) = m$ is a sufficient condition to estimate all unknown parameters. In the statistics literature, e.g. [Seb77], this property is known as identifiability. For linear regression, local and global identifiability coincide, because the Jacobian matrix, J , is independent of θ . Furthermore, there exist no local optima but exactly one global solution, θ , which minimises (4.3).

If $\text{Rank}(J) = m$ and $N > m$, an estimator of the variance of the measurement error, σ^2 , is given by: $s^2 = S(\hat{\theta})/(N - m)$. Notice here that in most practical cases the statistical properties are not known exactly, but assumed to have an error structure as in Section 1.6. The variance of the measurement error is not known. The following properties can be derived, [SW88], with \mathbf{E} denoting the expectation:

$$\mathbf{E}(s^2) = \mathbf{E}\left(\frac{S(\hat{\theta})}{N - m}\right) = \sigma^2, \quad (4.6)$$

$$\mathbf{E}(\hat{\theta}) = \theta^*, \quad (4.7)$$

$$\begin{aligned} \text{cov}(\hat{\theta}) &= \mathbf{E}\left((\hat{\theta} - \mathbf{E}(\hat{\theta}))(\hat{\theta} - \mathbf{E}(\hat{\theta}))^T\right) \\ &= \mathbf{E}\left((J^T J)^{-1} J^T Y Y^T J (J^T J)^{-1}\right) = \sigma^2 (J^T J)^{-1}. \end{aligned} \quad (4.8)$$

Which implies:

$$\hat{\theta} \sim \mathcal{N}(\theta^*, \sigma^2 (J^T J)^{-1}).$$

From (4.6) and (4.7), we see that the estimators for θ^* and σ^2 are unbiased in the linear case. Further, we need the following properties.

Theorem 4.2.1 Under the conditions $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\text{Rank}(J) = m$, the following properties hold:

$$1.) \quad \hat{\theta} - \theta^* \sim \mathcal{N}\left(0, \sigma^2 (J^T J)^{-1}\right);$$

$$2.) \quad S(\hat{\theta})/\sigma^2 \sim \chi^2(N - m);$$

$$3.) \quad \hat{\theta} \text{ is statistically independent of } s^2; \text{ and}$$

$$4.) \quad \frac{(S(\theta^*) - S(\hat{\theta}))/m}{S(\hat{\theta})/(N - m)} \sim \mathcal{F}(m, N - m), \quad (4.9)$$

where $\chi^2(N - m)$ and $\mathcal{F}(m, N - m)$ indicate the Chi-square distribution with $N - m$ degrees of freedom and Fisher's F-distribution with m and $N - m$ degrees of freedom, respectively.

Proof: See Seber and Wild [SW88, page 24]. \square

The $m \times m$ -matrix $\frac{1}{\sigma^2}(J^T J)$ is the so-called *Fisher information matrix*. A direct consequence of

$$S(\theta^*) - S(\hat{\theta}) = (\hat{\theta} - \theta^*)^T J^T J (\hat{\theta} - \theta^*) , \quad (4.10)$$

and (4.9) is:

$$\frac{(\hat{\theta} - \theta^*)^T J^T J (\hat{\theta} - \theta^*)}{ms^2} \sim \mathcal{F}(m, N - m) . \quad (4.11)$$

Consequently, a $(1 - \alpha)$ confidence region for θ^* is given by:

$$\{\theta^* : (\hat{\theta} - \theta^*)^T J^T J (\hat{\theta} - \theta^*) \leq ms^2 \mathcal{F}_\alpha(m, N - m)\} . \quad (4.12)$$

For a geometric interpretation of the ellipsoidal confidence region, we refer to the last paragraph of Section 1.6.

The remainder of this chapter is devoted to nonlinear regression. In the case of nonlinear regression the difference in (4.10) is not exact any more, but contains higher order terms, $\mathcal{O}(\|\hat{\theta} - \theta^*\|^3)$. This has consequences for the estimators and their confidence regions. Another main difference is the possibility of having many local minima in the nonlinear case. As a consequence, good initial estimates of the unknown parameters are indispensable to determine the optimal estimate efficiently.

4.3 Biased estimators

In the case of linear regression (4.7) holds, which means that $\hat{\theta}$ is an unbiased estimator of the true parameter vector, θ^* . In nonlinear regression the least squares estimator (LSE) is not always unbiased and the difference $\mathbf{E}(\hat{\theta}) - \theta^*$ is called the bias. To obtain insight in the meaning of the bias we start with an analytic computation of the bias. The cases where exact bias measures can be calculated analytically originate from carefully constructed examples and not from real life case studies, so we need other means to investigate the bias in a general setting. Besides the exact calculation of the bias we study two other methods, namely the Monte Carlo method and the bias measure of Box, to approximate the bias. Both these methods only yield approximate values for the bias, but they have the advantage that they are applicable in more general cases. In Sections 4.3.2 and 4.3.3 we discuss these methods and investigate their accuracy by applying them to the example introduced in Section 4.3.1.

4.3.1 Analytic result

In this section we look into the topic of bias by means of an exploratory example of a nonlinear regression problem. This example is constructed in such a way that the bias can be calculated analytically. The analytic result is compared with the approximate results of the following sections.

Example

We consider the model

$$y(t, \theta) = \ln(\theta + \ln(t)) , \quad (4.13)$$

where θ is the parameter to be estimated. For reasons which will become clear later, we take all N measurements at one fixed value: $\bar{t} > 0$. The additional parameter constraint reads: $\theta > -\ln(\bar{t})$. The simulated measurements are denoted by: $(1, \bar{t}, \tilde{y}_i)$, ($i = 1, \dots, N$) and the corresponding weights, w_i in (1.4), are taken equal. For convenience, the expectation of the measured values is scaled to 1, which means that $\tilde{y}_i = 1 + \varepsilon_i$ ($i = 1, \dots, N$), with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\theta^* = e - \ln(\bar{t})$.

The model of (4.13) and the chosen experimental design enable us to write an explicit expression for the optimal parameter:

$$\hat{\theta} = \exp(\bar{y}) - \ln(\bar{t}) , \quad \text{with } \bar{y} = \frac{1}{N} \sum_{i=1}^N \tilde{y}_i . \quad (4.14)$$

First, we get:

$$\hat{\theta} - \theta^* = e^{\bar{y}} - e , \quad (4.15)$$

and hence the bias of $\hat{\theta}$ equals:

$$\mathbf{E}(\hat{\theta} - \theta^*) = \mathbf{E}(e^{\bar{y}} - e) , \quad (4.16)$$

or, its complete expression:

$$\begin{aligned} \mathbf{E}(\hat{\theta} - \theta^*) &= \int_{-\infty}^{\infty} (e^{\bar{y}} - e) \sqrt{\frac{N}{2\pi\sigma^2}} \exp\left(-\frac{N(\bar{y} - 1)^2}{2\sigma^2}\right) d\bar{y} \\ &= \int_{-\infty}^{\infty} \sqrt{\frac{N}{2\pi\sigma^2}} \exp\left(\bar{y} - \frac{N(\bar{y} - 1)^2}{2\sigma^2}\right) d\bar{y} - e \\ &= \int_{-\infty}^{\infty} \sqrt{\frac{N}{2\pi\sigma^2}} \exp\left(-\frac{N\left\{\bar{y} - \left(1 + \frac{\sigma^2}{N}\right)\right\}^2}{2\sigma^2}\right) \exp\left(1 + \frac{\sigma^2}{2N}\right) d\bar{y} - e \\ &= e \left(\exp\left(\frac{\sigma^2}{2N}\right) - 1 \right) . \end{aligned} \quad (4.17)$$

Therefore, in this example $\hat{\theta}$ is a biased estimate of θ^* . The magnitude of this bias is shown in Table 4.1 for different values of σ^2/N . \diamond

4.3.2 Monte Carlo

The purpose of this section is to motivate and to explain our Monte Carlo (MC) method. The method is demonstrated by making use of the example of Section 4.3.1. The MC-result is compared with the exact result from (4.17).

The method is used here to approximate the bias of an estimator. The bias can only be calculated analytically if an explicit relation between the estimator and the measurements exists as in (4.14). This is more of an exception than a rule, so we need alternative ways to approximate the bias.

Before the MC-method can start we need an experimental design, $\{c_i, t_i\}$, and an estimate of the unknown parameters, $\hat{\theta}$, and an estimate of the variance², s^2 . Then we perform repetitive perturbations of the model outputs $y_{c_i}(t_i, \hat{\theta})$, ($i = 1, \dots, N$) and repeat this N_{MC} times. In the case the measurement errors are independent and normally distributed, the perturbations are sampled from the same distribution. Each set of N artificial, simulated measurements has a corresponding least squares estimate (LSE). In the case the model is linear in its parameters, the N_{MC} corresponding LSEs will also have a normal distribution (see Theorem 4.2.1). In the nonlinear case, normality tests, e.g. via sample moments or the Kolmogorov-Smirnov test, on these LSEs give an indication of the nonlinearity of the regression problem and should be compared with nonlinearity information obtained by means of other methods. If we use the estimates $\hat{\theta}$ and s^2 for θ^* and σ^2 , respectively, the i -th artificial measurement of the j -th MC simulation reads:

$$\tilde{y}_i^j = f_{c_i}(t_i, \hat{\theta}) + \delta_i^j \quad (i = 1, \dots, N, \quad j = 1, \dots, N_{MC}), \quad (4.18)$$

with:

$$\delta_i^j \sim \mathcal{N}(0, s^2) .$$

In statistics literature this MC procedure is called parametric bootstrap [Efr79]. Every set of simulated measurements leads to a corresponding least squares estimate, denoted by: $\hat{\theta}^j$. The mean of these N_{MC} estimates is denoted by $\hat{\bar{\theta}}$. The difference between $\hat{\bar{\theta}}$ and $\hat{\theta}$ is the bootstrap estimate of the bias. The accuracy of the corresponding estimator depends on –besides the model and the experimental design– the number of estimates, N_{MC} . As an approximate $(1 - \alpha)$ -confidence region for $\hat{\theta}$ is given by (4.12), the same relation, given the estimate, can be used for $\hat{\bar{\theta}}$, with s^2 replaced by s^2/N_{MC} . When we perform, for instance, N_{MC} runs, the individual confidence regions of the bias are $\sqrt[3]{N_{MC}}$ times smaller than the individual confidence regions of the final estimate. This seems accurate enough for the bias, but this is not true. First, for the bias we consider the difference, $\hat{\bar{\theta}} - \hat{\theta}$, given the estimate, $\hat{\theta}$. Second, the bias reveals information with respect to the nonlinearity of the parameter estimation problem, because the bias is used for another purpose than the final estimates, it requires a different accuracy.

²Of course we use θ^* , $\mathbf{E}(y_{c_i}(t_i, \theta^*))$ or σ^2 , if these quantities are known.

If we return to the example of Section 4.3.1 the conditional variance for $\hat{\theta}$ given $\hat{\theta}$ can be estimated by:

$$\text{var}(\hat{\theta}) = (J^T J)^{-1} \frac{s^2}{N_{MC}} = \frac{e^2 s^2}{N_{MC} N},$$

because θ^* is known we take the true parameter instead of its estimate, $\hat{\theta}$. The results of the MC-method, the corresponding N_{MC} 's and the comparisons with the analytic results of (4.17) are shown in Table 4.1 for various ratios of σ^2 and N .

σ^2/N	anal. (4.17)	N_{MC}	$\hat{\theta} - \theta^*$
1.0×10^0	1.763×10^0	1.0×10^3	1.718×10^0
1.0×10^{-1}	1.394×10^{-1}	1.5×10^4	1.391×10^{-1}
1.0×10^{-2}	1.363×10^{-2}	1.6×10^5	1.392×10^{-2}
1.0×10^{-3}	1.359×10^{-3}	1.6×10^6	1.326×10^{-3}

Table 4.1: Bias estimates for the model problem of (4.13), calculated by analytic means, cf. (4.17) and the MC-method.

The MC-results are in close correspondence with the analytic results, although to our experience many simulations had to be performed to obtain accurate approximations. The choice of N_{MC} is made in such a way that the relative error between the true and the estimated bias is less than 5%. The number of MC runs might become a serious bottleneck for more complex models due to huge CPU times for model evaluations. If this is the case, we can approximate the bias as outlined in the next section.

4.3.3 Bias measure of Box

A useful bias measure was introduced by Box in [Box71]. We only give the formula of this bias measure, for details and the derivation the reader is referred to the original paper. For this bias measure we need the Jacobian, J (cf. (1.11)), and the Hessian, H (cf. (1.17)). The bias measure according to Box, abbreviated by BB , is defined by:

$$BB(\hat{\theta}) = \frac{-\sigma^2}{2} (J^T J)^{-1} J^T z, \quad (4.19)$$

where z is the N -dimensional vector:

$$z = \left(\text{Tr} \left(H_{1..} (J^T J)^{-1} \right), \text{Tr} \left(H_{2..} (J^T J)^{-1} \right), \dots, \text{Tr} \left(H_{N..} (J^T J)^{-1} \right) \right)^T,$$

and $H_{i..}$, the i -th *site* of H , is an $m \times m$ matrix. In (4.19) the matrices J and H are evaluated at $\hat{\theta}$. From (4.19) it is obvious that $BB(\hat{\theta})$ vanishes for linear models. When we calculate $BB(\hat{\theta})$ for the example from Section 4.3.1, we obtain:

$$BB(\hat{\theta}) = \frac{\sigma^2}{2N} \left(\hat{\theta} + \ln(\bar{t}) \right) . \quad (4.20)$$

Another way to look at (4.20) is to substitute the true parameter value, $\theta^* = e - \ln(\bar{t})$, for $\hat{\theta}$. This substitution yields:

$$BB(\theta^*) = \frac{e\sigma^2}{2N} , \quad (4.21)$$

which is a first term of the Taylor expansion of (4.17). The results are listed in Table 4.2. The expectation of (4.20) reads:

$$\mathbf{E} \left(BB(\hat{\theta}) \right) = \mathbf{E} \left(\frac{\sigma^2}{2N} (\hat{\theta} - \theta^* + e) \right) = \frac{\sigma^2}{2N} \left(\mathbf{E}(\hat{\theta} - \theta^*) + e \right) , \quad (4.22)$$

which gives the relation between the true bias and the bias measure of Box. The values of the expected bias measure of Box, using the exact biases, are given in Table 4.2. This table indicates that in this example the quadratic approximation of the bias measure of Box is acceptable, if σ^2/N is an order of magnitude smaller than e .

σ^2/N	anal. (4.17)	$BB(\theta^*)$ (4.21)	$\mathbf{E} \left(BB(\hat{\theta}) \right)$
1.0×10^0	1.763×10^0	1.359×10^0	2.241×10^0
1.0×10^{-1}	1.394×10^{-1}	1.359×10^{-1}	1.429×10^{-1}
1.0×10^{-2}	1.363×10^{-2}	1.359×10^{-2}	1.366×10^{-2}
1.0×10^{-3}	1.359×10^{-3}	1.359×10^{-3}	1.360×10^{-3}

Table 4.2: Bias measures of Box (4.19) for the model problem (4.13).

4.4 Curvature measures

The bias measures as they have been derived in the previous sections give only a limited amount of information about the nonlinearity. When they indicate that the bias is negligible, we do not need additional information to proceed the investigation of the nonlinearity. If this is not the case, we want to explore the nonlinearity in more detail. First, we give a short overview of the curvature measures proposed by Bates and Watts [BW88], we highlight the problems which might be encountered in nonlinear regression, and we show how to recognise them and to deal with them.

It is important to keep in mind that the expression ‘measure of nonlinearity’ can be misleading when only second order information is used. Although we will follow the literature here, it would be better to call the existing measures: measures of quadraticity. A model that is cubic in its parameters, could be called linear according to the measures of Bates and Watts.

To get more insight into the essential differences between linear and nonlinear regression and in order to describe measures for nonlinearity we have to introduce the notion of *solution locus*. Each set of N measurements can be regarded as one point in an N -dimensional data space. The solution locus, is the m -dimensional manifold in the data space, containing all possible, theoretical model responses for all possible θ . In the case the dimension of the solution locus is (locally) less than m , the problem is (locally) non-identifiable, for more details on identifiability the reader is referred to [WP97]. The orthogonal projection of the point, which corresponds to the actual measurements, onto the solution locus leads to the LSE, $\hat{\theta}$. Notice that the solution locus does not depend on the measurements, $\{\tilde{y}_i\}$, but only on the model outcome, $y(t, \theta)$, and the experimental design, $\{c_i\}$ and $\{t_i\}$. The nonlinearity of the model-experiment combination can be expressed in terms of the curvature of the solution locus.

Let us first give an example in order to illustrate the solution locus.

Example

Suppose that we have a chemical reaction where two substances, A and B , are involved, and the reaction scheme is given by:



When we assume first order reaction kinetics and the reaction starts at t_0 , the differential equation describing this chemical reaction reads:

$$\frac{d[A]}{dt} = -k[A] , \quad \text{scaling: } [A] \text{ such that: } [A](t_0) = [A]_0 = 1 ,$$

we obtain $[A](t) = e^{-kt}$. From now on the unknown parameter k is written as $\theta := k \geq 0$.

We assume that two measurements have been performed at $t = 1$ and $t = 2$. Using the notation (1.2), $N = 2$ and the experimental data are given by $\{(1, 1, \tilde{y}_1), (1, 2, \tilde{y}_2)\}$. We have a two-dimensional data space and a one-dimensional solution locus, given by the parametric form:

$$\begin{pmatrix} y_{c_1}(t_1, \theta) \\ y_{c_2}(t_2, \theta) \end{pmatrix} = \begin{pmatrix} e^{-\theta} \\ e^{-2\theta} \end{pmatrix} , \quad \text{with: } \theta \geq 0 . \quad (4.23)$$

The data space containing the solution locus is shown in Figure 4.1. Thus, the solution locus contains all theoretical model responses for the given experimental design. Each complete set of N measurements corresponds to a single point in the data space. For a given set of experimental data the LSE is determined by orthogonal projection onto the

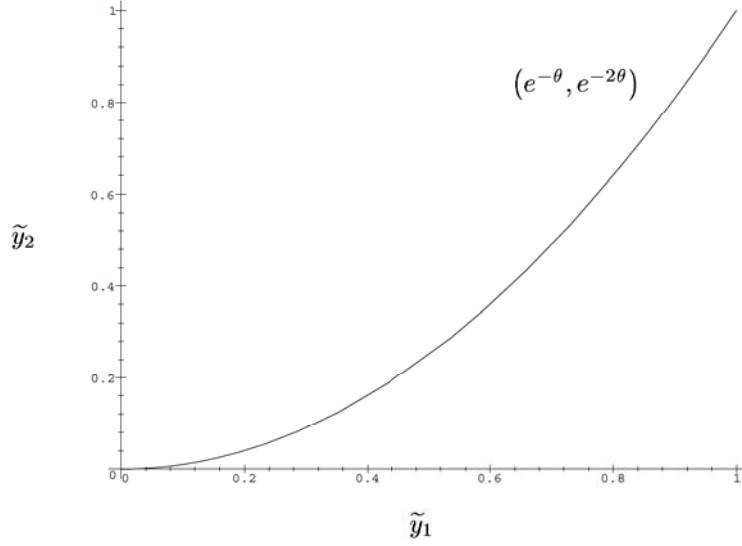


Figure 4.1: Plot of data space and solution locus, cf. (4.23).

solution locus. If the weights in (1.4) are not constant, then the axes of the data space should be scaled by the corresponding weights. In our case, taking $w_1 = w_2 = 1$, the sum of squared discrepancies reads:

$$S(\theta) = (e^{-\theta} - \tilde{y}_1)^2 + (e^{-2\theta} - \tilde{y}_2)^2 = Y^T(\theta)Y(\theta) . \quad (4.24)$$

and the LSE, $\hat{\theta}$, can be obtained by an orthogonal projection. The discrepancy vector, $Y(\theta)$, connects the measurements to the solution locus, and $\partial Y(\theta)/\partial \theta$, is the tangent plane of the solution locus. From (4.4) we see that these two quantities are orthogonal, if $S(\theta)$ has a vanishing gradient.

Taking the derivative of (4.24) with respect to θ and setting it equal to zero leads to the implicit equation describing $\hat{\theta}$ as a function of \tilde{y}_1 and \tilde{y}_2 :

$$e^{-3\hat{\theta}} + (\frac{1}{2} - \tilde{y}_2)e^{-\hat{\theta}} - \frac{1}{2}\tilde{y}_1 = 0 . \quad (4.25)$$

The surface representing this relation is given in Figure 4.2

◇

After the example we return to the general notation. The solution locus, (cf. (4.23)), is now denoted by:

$$\left\{ \eta(\theta) = (\eta_1(\theta), \dots, \eta_N(\theta))^T \mid \theta \in \mathbb{R}^m \right\} , \quad (4.26)$$

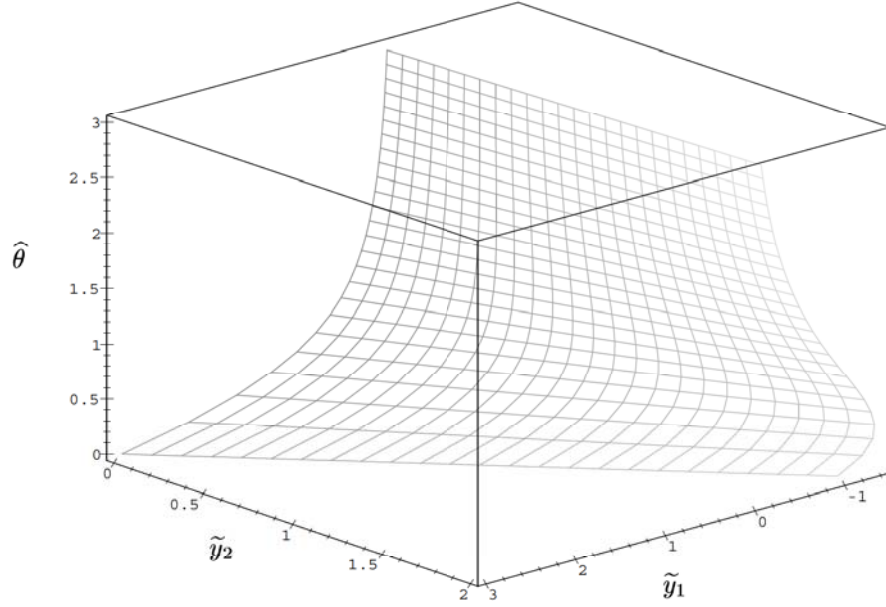


Figure 4.2: Plot of the surface describing the least squares estimate, $\hat{\theta}$, by a given pair of measurement, \tilde{y}_1 and \tilde{y}_2 , as expressed by (4.25).

where

$$\eta_i(\theta) = y_{c_i}(t_i, \theta) , \quad \text{for: } i = 1, \dots, N .$$

Obviously, linearity of the model leads to a linear solution locus. In order to get an impression of the nonlinearity of the solution locus we consider an arbitrary straight line in the parameter space through $\hat{\theta}$, denoted by:

$$\theta(\beta) = \hat{\theta} + \beta h , \quad 0 \neq h \in \mathbb{R}^m, \beta \in \mathbb{R} .$$

The model transforms this straight line into a curved line on the solution locus:

$$\eta_h(\beta) = \eta(\hat{\theta} + \beta h) .$$

In literature this curve is called the *lifted line*. This is a straight line if the model is linear in θ . The tangent to the lifted line at $\hat{\theta}$ is given by

$$\left. \frac{d\eta_h(\beta)}{d\beta} \right|_{\beta=0} = \dot{\eta}_h(0) = J(\hat{\theta})h .$$

This means that the columns of the Jacobian matrix, J , span the tangent plane of the solution locus. The second derivative (the ‘acceleration of a particle travelling along the lifted line’):

$$\ddot{\eta}_h(0) = h^T H h ,$$

is an N -dimensional vector, which can be split into two parts, $\ddot{\eta} = \ddot{\eta}^\perp + \ddot{\eta}^\parallel$. One part, denoted by $\ddot{\eta}^\perp$, corresponds to the acceleration perpendicular to the tangent plane. The other part, $\ddot{\eta}^\parallel$, denotes the acceleration in the tangent plane. From these second derivatives we compute the curvatures of the solution locus. The *normal curvature* in direction h is defined as:

$$K_h^\perp = \frac{\|\ddot{\eta}_h^\perp\|}{\|\dot{\eta}_h\|^2} . \quad (4.27)$$

This normal curvature equals the reciprocal of the radius of the circle which osculates the solution locus in the direction of $\dot{\eta}_h$ at $\eta(\hat{\theta})$. This curvature measure is a characteristic of the solution locus, determined by the model $y(t, \theta)$, the choice of h , and the experimental design $\{c_i, x_i\}$.

The curvature derived from the tangential acceleration

$$K_h^\parallel = \frac{\|\ddot{\eta}_h^\parallel\|}{\|\dot{\eta}_h\|^2} , \quad (4.28)$$

is called the *parameter-effect curvature* in the direction h .

Before we explore the meaning of these curvatures, we make them scale invariant. Because multiplication of the model responses by a factor, say κ , leads to a curvature which is $1/\kappa$ times the original one, the curvatures are scaled by the *standard radius* (cf. [BW88, page 242]),

$$\rho = \sqrt{\frac{m}{N-m} S(\hat{\theta})} = s\sqrt{m} . \quad (4.29)$$

Notice that ρ^2 is also used in the denominator of (4.12). This standard radius depends on $S(\hat{\theta})$ and decreases if the model fits the data better. The relative (scale invariant) curvatures are defined by:

$$\gamma_h^\perp = K_h^\perp \rho \quad \text{and} \quad \gamma_h^\parallel = K_h^\parallel \rho . \quad (4.30)$$

The *relative normal curvature* is a measure for the deviation between the solution locus and its tangent plane. The $(1 - \alpha)$ confidence region from (4.12) is a disc with radius $\rho\sqrt{\mathcal{F}_\alpha(m, N-m)}$ on the tangent plane, centred at $\eta(\hat{\theta})$. If the radius of the smallest circle osculating the solution locus is at least twice as big as the radius of the $(1 - \alpha)$ confidence region, i.e.: $\gamma_h^\perp < 1/(2\sqrt{\mathcal{F}_\alpha(m, N-m)})$ for all directions h , then the relative deviation between the tangent plane to the osculating circle is less than 13.4%. The

planar assumption is very likely if this inequality holds. The *relative parameter-effect curvature* measures the distortion of a rectangular grid in the parameter space into a non-rectangular grid on the solution locus due to the mapping of (4.26).

From now on we always use the relative curvature measures, γ_h , and therefore omit the adjective relative. The parameter-effect curvature can, contrary to the normal curvature, be decreased by an appropriate reparametrisation. For this reason the normal curvature is also known as the *intrinsic curvature*. Further we want to emphasize that a study of the parameter-effect curvature and a possibly appropriate reparametrisation are only constructive if the intrinsic curvature is sufficiently small.

In order to calculate both curvatures we consider the QR -decomposition of the Jacobian matrix:

$$J = QR = Q \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix},$$

where Q is an orthonormal, $N \times N$ -matrix and \tilde{R} is an upper triangular, $m \times m$ -matrix. The matrix \tilde{R} is used for a linear coordinate transformation in the parameter space

$$\phi := \tilde{R}(\theta - \hat{\theta}).$$

Notice, that a linear transformation will not affect the measures of nonlinearity, so it makes no difference whether we study the nonlinearity measures with respect to θ or ϕ . Here we assume that $\text{Rank}(J) = m$ in a vicinity of $\hat{\theta}$, which means that the problem is locally identifiable. The consequences for the case $\text{Rank}(J) < m$ are discussed in Section 4.6. Consequently, the inverse of \tilde{R} exists. When we now consider the derivatives of η with respect to ϕ and denote the corresponding Jacobian by J_ϕ , we get:

$$J_\phi = \left. \frac{d\eta}{d\phi} \right|_{\phi=0} = \left. \frac{d\eta}{d\theta} \right|_{\theta=\hat{\theta}} \left. \frac{d\theta}{d\phi} \right|_{\phi=0} = Q \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} \tilde{R}^{-1} = Q \begin{pmatrix} I_m \\ 0 \end{pmatrix}. \quad (4.31)$$

This means that the first m columns of Q contain an orthonormal basis of the tangent plane, $d\eta(\phi)/d\phi$. If we denote \tilde{R}^{-1} by L , the second derivatives of η with respect to ϕ , i.e. the Hessian after the linear transformation, turn into:

$$(H_\phi)_{ijk} = \frac{d^2 \eta_i(\phi)}{d\phi_j d\phi_k} = \sum_{q=1}^m \sum_{p=1}^m \frac{d^2 \eta_i(\theta)}{d\theta_p d\theta_q} \frac{d\theta_p}{d\phi_j} \frac{d\theta_q}{d\phi_k}, \quad (4.32)$$

or using a notation with the sites of H as introduced in (4.19):

$$(H_\phi)_{i..} = L^T H_{i..} L. \quad (4.33)$$

Now we are going to split the $m \times m$ vectors of length N with second derivatives into a tangent and a normal part. Therefore we multiply this matrix, H_ϕ , by Q^T :

$$(A)_{ijk} = \sum_{l=1}^N Q_{il}^T (H_\phi)_{ljk}. \quad (4.34)$$

The ‘upper’ part of A , A_{ijk} for $i, j, k = 1, \dots, m$, also called the first m sites of A , contain entries with respect to parameter-effect curvature and the last $N-m$ sites contain intrinsic curvature information, denoted by A^\parallel and A^\perp , respectively. An advantage of the transformation becomes clear if we take a vector, say g , from the rotated parameter space in such a way that $\|g\| = 1$, then $\|\dot{\eta}_{Lg}\|$ also equals 1. And therefore we obtain, by using (4.30), (4.33) and (4.34):

$$\gamma_{Lg}^\parallel = \|(g^T H_\phi g)^\parallel\| \rho = \|g^T A^\parallel g\| \rho \quad (4.35)$$

and

$$\gamma_{Lg}^\perp = \|(g^T H_\phi g)^\perp\| \rho = \|g^T A^\perp g\| \rho. \quad (4.36)$$

From the two relative curvatures, we denote the corresponding maxima as:

$$\begin{aligned} \Gamma^\perp &= \max_{\|g\|=1} \gamma_{Lg}^\perp, \\ \Gamma^\parallel &= \max_{\|g\|=1} \gamma_{Lg}^\parallel, \end{aligned}$$

and the corresponding vector in the rotated parameter space by \hat{g}^\perp and \hat{g}^\parallel , respectively. An algorithm for this maximisation is proposed in [BW80].

If both Γ^\perp and Γ^\parallel do not exceed $1/(2\sqrt{\mathcal{F}_\alpha(m, N-m)})$ the nonlinearity of the parameter estimation problem is marginal and the linear theory can be applied. To be sure at this point it is still recommended to compare these results with other measures. In the case Γ^\perp is too large, then CPU intensive methods are needed to sample in the vicinity of $\hat{\theta}$ to retrieve confidence regions. Another option is when Γ^\parallel exceeds the corresponding F-value, then a reparametrisation might give some decrease of the nonlinearity. For this purpose we transform \hat{g}^\parallel linearly from the rotated parameter space into the original θ -space: $\hat{h}^\parallel = L\hat{g}^\parallel$. The entries of \hat{h}^\parallel which differ substantially from zero, indicate that the corresponding parameters should be considered for a reparametrisation. The choice of the reparametrisation depends on the experience and intuition of the user, the nonlinearity measures indicate only which parameters are candidates for a reparametrisation in order to reduce the parameter-effect curvature. Examples of a successful reparametrisation are given in the example below and, for a practical case study, in Section 6.1.

Example

We return to the example described by (4.23) and assume that the measurements are known, say: $\tilde{y}_1 = 0.61$ and $\tilde{y}_2 = 0.46$. Equation (4.25) leads to $\hat{\theta} = 0.411$. These measurements are also needed to scale the curvatures. Substitution in (4.29) leads to: $\rho = 0.0532$. The scaled Jacobian and Hessian are

$$\begin{aligned} J &= \frac{1}{0.0532} \begin{pmatrix} -e^{-0.411} \\ -2e^{-0.821} \end{pmatrix} = \begin{pmatrix} -12.46 \\ -16.53 \end{pmatrix}, \\ H &= \frac{1}{0.0532} \begin{pmatrix} e^{-0.411} \\ 4e^{-0.821} \end{pmatrix} = \begin{pmatrix} 12.46 \\ 33.06 \end{pmatrix}. \end{aligned}$$

Subsequently we compute the QR decomposition of the Jacobian:

$$J = QR = Q \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} = \begin{pmatrix} -0.602 & 0.799 \\ -0.799 & -0.602 \end{pmatrix} \begin{pmatrix} 20.70 \\ 0 \end{pmatrix}.$$

Because $L = \tilde{R}^{-1} = 0.0483$, the Hessian with respect to the transformed parameters reads:

$$H_\phi = L^T H L = 0.0483 \begin{pmatrix} 12.46 \\ 33.06 \end{pmatrix} 0.0483 = \begin{pmatrix} 0.0291 \\ 0.0772 \end{pmatrix}$$

Finally, we get the matrix which contains both curvatures:

$$A = Q^T H_\phi = \begin{pmatrix} -0.0791 \\ 0.0232 \end{pmatrix}$$

The absolute value of the first and the second entry of this matrix correspond with γ^\parallel and γ^\perp , respectively. Note that $1/(2\sqrt{\mathcal{F}_{0.05}(1,1)}) = 0.0394$, which means that there is a strong nonlinearity in the parameter-effect part.

If we use the reparametrisation:

$$\psi = e^{-\theta}, \quad (4.37)$$

the corresponding acceleration array reads:

$$A_\psi = \begin{pmatrix} 0.0308 \\ -0.0232 \end{pmatrix}$$

We see that the parameter-effect curvature decreases (which was the aim). The intrinsic curvature stays unaffected as expected from the theory. \diamond

In the case that ρ becomes larger, the intrinsic and the parameter-effect curvature will also increase due to (4.30). The quantity ρ is introduced to scale the error, which is dependent on neither the model nor the experimental design. When we have a look at Figure 4.1, it is obvious that if $\hat{\theta}$ increases (and ρ is kept constant), the curvatures also increase.

The normal (non-relative) curvature (4.27) corresponds with the radius of the circle which osculates the solution locus. If the measurements coincide with the centre of the osculation circle, the problem becomes locally non-identifiable. This can happen even if $\gamma_h^\perp < 1/(2\sqrt{\mathcal{F}_\alpha(m, N-m)})$.

4.5 Investigation of levelsets

In this section we give a collection of guidelines which are valuable to investigate the significance of the ellipsoidal confidence region (4.12) based on a linear approximation. The

guidelines have a heuristic character, but contribute in our point of view to get a better insight into the nonlinearity of a regression problem. The extent of the correspondence between the approximated levelsets and the true levelsets is related to the nonlinearity of the regression problem. The guidelines vary from retrieving rough information about this correspondence in a cheap way up to more sophisticated and time consuming approaches to investigate the levelsets more precisely. Information about the nonlinearity from other sources can be integrated with these guidelines. The sum of squared discrepancies for an ellipsoidal $(1 - \alpha)$ -levelset is denoted by S^α and equals:

$$S^\alpha = S(\hat{\theta}) \left(1 + \frac{m}{N-m} \mathcal{F}_\alpha(m, N-m) \right).$$

For a first exploration we compute the sum of squared discrepancies at the intersections of the ellipsoid with the parameter axes (see (1.26)) and compare the corresponding sums of squared discrepancies with the value S^α for different values of α . This can be repeated for the tips of the ellipsoid. For each confidence level we obtain $4m$ sums of squares, denoted by $S^{\alpha,i}$, $i = 1, \dots, 4m$. The deviation from linearity can be expressed by

$$\mu^\alpha = \frac{\sqrt{\sum_{i=1}^{4m} (S^\alpha - S^{\alpha,i})^2}}{2\sqrt{m}S(\hat{\theta})},$$

which is scale invariant, corrected for the number of points on the ellipse and zero in the linear case.

Instead of taking only $4m$ points at the intersections and the tips of ellipsoid, we can take an arbitrary number of points on the ellipse and calculate the corresponding μ^α . The points can be either sampled randomly on the ellipse or positioned on the ellipse in a regular way. The computation of such a regular positioning on a sphere is discussed in [PSS97], the extension to an ellipse is straightforward.

Starting from N_α points on the ellipse, denoted by $\theta^{\alpha,i}$, we can perform a line search along the line through $\hat{\theta}$ and $\theta^{\alpha,i}$, in order to retrieve $\check{\theta}^{\alpha,i}$, s.t. $S(\check{\theta}^{\alpha,i}) = S^\alpha$. The resulting points, $\check{\theta}^{\alpha,i}$ ($i = 1, \dots, N_\alpha$), should be projected on all $\{\theta_i, \theta_j\}$ -planes ($1 \leq i < j \leq m$) and compared with the corresponding, projected ellipse. Similar to μ^α , we can derive another heuristic measure of nonlinearity:

$$\omega^\alpha = \frac{\sqrt{\sum_{i=1}^{N_\alpha} \|\theta^{\alpha,i} - \check{\theta}^{\alpha,i}\|^2}}{\sqrt{N_\alpha} \|\hat{\theta}\|},$$

which is, as μ^α , scale invariant, corrected for the number of points on the ellipse and zero in the linear case.

A straightforward approach is to use a grid in the parameter space around $\hat{\theta}$, calculate the corresponding sum of squares, make iso-plots of all the $\binom{m}{2}$ intersections with the $\{\theta_i, \theta_j\}$ -planes and compare the results with the ellipsoidal regions which were expected

on the basis of the linear theory. The disadvantage is that the computation time grows exponentially with m , although a priori knowledge about the nonlinearity of certain parameters can be used to refine the grid in the direction of these parameters. An example of an iso-plot and the comparison with an ellipsoidal region is given in Section 6.1.7.

For all the methods it is important to keep in mind that for the purpose of visualisation not only the intersections with $\{\theta_i, \theta_j\}$ -planes ($1 \leq i < j \leq m$) should be considered, but also the projection on such planes. To demonstrate the last sentence we can think of a banana-shaped levelset whose intersections with the $\{\theta_i, \theta_j\}$ -planes are almost ellipsoids and only the projection will reveal the banana-shape of the levelset. For this reason it also not recommended to sample points in $\{\theta_i, \theta_j\}$ -planes only, because sampling points over the whole parameter space might reveal additional information. Here it is important to remark that when m grows the projected points of a more dimensional ellipse concentrate more at the centre of the projected ellipse. This is a disadvantage as long as we are interested in the contours of levelsets and their graphical representation.

4.6 Parameter constraints and redundancy

In this section we will give a short outline concerning active parameter constraints and the consequences for the confidence region. At the end of this section we highlight a few topics with respect to parameter redundancy, which is related to over-parametrisation and non-identifiability.

The $(1 - \alpha)$ -confidence region indicates the area that has probability $(1 - \alpha)$ to cover the true parameter, θ^* . In this section we will assume implicitly that θ^* fulfills the constraints (1.27), i.e.: $R(\theta^*) \leq 0$. In the case that none of the $(m - 1)$ -dimensional manifolds $R_i(\theta) = 0$ ($i = 1, \dots, K$) intersects the confidence region, this confidence region stays unchanged.

In the case when there is such an intersection a number of steps have to be made. First, we concentrate on the physical interpretation of this situation. E.g., when a reaction rate tends to zero, we have to perform statistical tests in order to decide whether this reaction is insignificant, and as a consequence the corresponding parameter and restriction can be omitted. Then the model is adjusted and fit to the data again. Second, if it turns out that a restriction intersecting the confidence region does not have such a consequence in the proper formulation of the model, then the area

$$\Theta_1 = \{\theta | S(\theta) \leq S^\alpha \wedge R(\theta) \leq 0\}$$

still has a probability of $(1 - \alpha)$ that it covers θ^* .

When a parameter estimation problem is non-identifiable in the linear case we have

$$\text{Rank}(J) < m \Leftrightarrow \exists \delta\theta \neq 0 | S(\hat{\theta} + \delta\theta) = S(\hat{\theta}),$$

where $\delta\theta \in \text{Ker}(J)$ and $\hat{\theta}$ is a non-unique point in the parameter space which minimises $S(\theta)$. The rank of J , denoted by m_j , determines how many parameters can be estimated from the parameter estimation problem.

When J is singular we can still retrieve the corresponding singular value decomposition (cf. (1.15)) of J , such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{m_J} > 0$ and $\lambda_{m_J+1} = \dots = \lambda_m = 0$. The last $m - m_J$ columns of V span the kernel of J . The parameter transformation

$$\phi = V^T(\theta - \hat{\theta}) \quad (4.38)$$

leads to

$$J_\theta = \frac{dY(\theta)}{d\theta} = \frac{dY(\phi)}{d\phi} \frac{d\phi}{d\theta} = J_\phi V^T,$$

and as a result:

$$J_\phi^T J_\phi = \Sigma^2.$$

After the parameter transformation of (4.38), the parameters $\phi_{m_J+1}, \dots, \phi_m$ can be deleted from the model equations. The remaining parameters are called the *principal components*, the corresponding Jacobian has full rank and the parameters are uncorrelated.

In practical situations the true rank of a matrix is not an appropriate measure due to expected numerical truncation errors. Therefore, we consider the ϵ -rank or ‘numerical rank’ of a matrix, see [GV83, page 176]. This ϵ -rank of $\hat{J}(\theta)$, $m_{\epsilon,J}$, equals the largest i such that $\lambda_i > \epsilon \lambda_1$. For parameter estimation problems a choice of ϵ between 10^{-3} and 10^{-5} is sufficient.

If both γ_h^\perp and γ_h^\parallel are smaller than $1/(2\sqrt{\mathcal{F}_\alpha(m, N-m)})$, the regression problem is assumed to be close to linear and the linear approximation for the level sets is assumed to be valid. To be more sure we check whether this quadratic information is in accordance with heuristic techniques from Section 4.5. If this check is positive we can perform the parameter transformation (4.38) in the vicinity of $\hat{\theta}$ for the mentioned values of ϵ .

4.7 Concluding remarks

In this chapter we started with a brief overview of linear regression, which was followed by a summary of the differences between linear and nonlinear regression. Special attention was paid to ways to quantify the nonlinearity of a regression problem. Some of the approaches to derive nonlinearity measures require a huge amount of model evaluations, which make them less appropriate in the case the model equations consist of a set of DAEs.

The nonlinearity measures can be used to obtain a clue with respect to a reparametrisation or an educated sample strategy in the parameter space. Various aspects of nonlinear regression are illustrated by examples in this chapter or related to the case studies of Chapter 6.

Chapter 5

Optimal Experiment Design

5.1 Introduction

In the previous chapters we have focused on parameter estimation, model discrimination and the corresponding statistical analyses, all on the basis of a given, fixed set of measurements. If the results from the statistical analyses are insufficient to discriminate between two models or give rise to large, unwanted confidence regions of the parameters, we need additional measurements in order to obtain a decisive answer or more precise estimates. A third goal for future experiments could be the reduction of the nonlinearity of a regression problem. Except for a simple example we will not deal with this topic, although it is a promising and targeting topic for future research.

Parameter estimation is an initial step towards a more thorough investigation of the model. Optimal experiment design studies the issue of how to plan future experiments in order to obtain a maximum of information. The kind of information depends on the motivation of a more thorough investigation and is specified mathematically in this chapter.

Section 5.2 directs to a more precise, mathematical formulation of the topic. A method to get a clear insight at a glance into the dependencies between the state variables and the parameters is presented in Section 5.3. The Sections 5.4 and 5.5 contain an outline on optimal experiment design (OED) in order to reduce the size of the confidence regions and to discriminate between different models, respectively. A relation with nonlinearity is given in Section 5.6. Concluding remarks are found in Section 5.7.

Again the model responses are denoted by $y_{c_i}(t_i, \theta)$, where the pairs $\{c_i, t_i\}$ ($i = 1, \dots, N$) specify the experimental design. As before in Chapter 4, we assume that the Jacobian and the Hessian –or their numerical approximations– exist for the given experimental design.

5.2 Problem formulation

In the first paragraphs of this section we focus on the problem formulation for the case that a model has been selected and we want to reduce the size of the confidence region for the parameters. The last part of this section is devoted to a more precise formulation

of optimal experiment design for model discrimination purposes.

Besides the model equations, we assume the presence of either a good estimate for the unknown parameters or a set of measurements that can be used to estimate these unknowns. It is more a rule than an exception that some entries of θ cannot be estimated with acceptable reliability. In such cases it is of major interest to put effort in the design of future experiments in order to reduce the uncertainty in the estimators of these parameters. A schematic flow chart of a model investigation and the position of optimal experiment design is given in Figure 5.1.

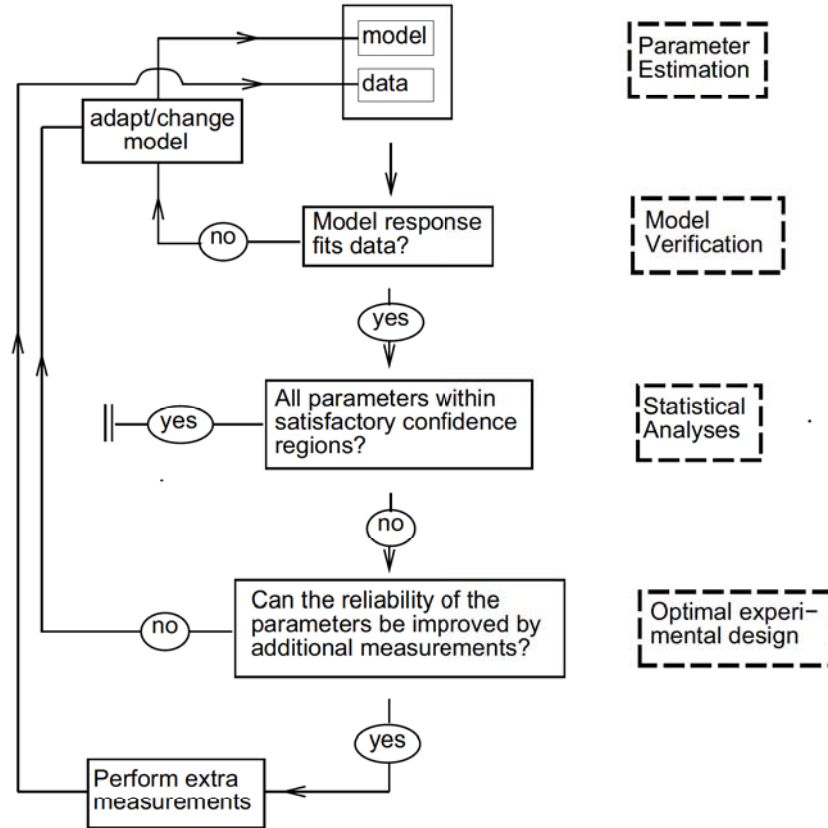


Figure 5.1: Schematic representation of a model investigation, where optimal experiment design is used to increase the reliability of the estimators of the parameters.

Here we encounter one of the motivations for optimal experiment design. Given a set of parameters and the reliability of the corresponding estimators, what additional

experiments should be performed to increase this reliability in a well defined sense? This is a bit of a paradoxical task in the nonlinear case, because the parameters with their uncertainty also influence the optimal experiment design. Consequently, the optimal experiment design is based on the current estimates and might turn out to be far from optimal when the estimates change after having performed additional measurements.

The sum of squares to be minimised and its $N \times m$ Jacobian are denoted as in (1.10) and (1.11), respectively. In the case $N < m$ the approach as it will be presented in the following sections is still applicable, $\hat{\theta}$ is then one of all possible least squares estimates or an estimate based on other information.

If the nonlinearity measures are sufficiently small (cf. Section 4.4), then the ellipsoidal region of (1.24), which is only a linear approximation, shows close correspondence with the true confidence region. Therefore, investigating $J(\hat{\theta})$ yields a reliable basis to retrieve an optimal experiment design with the aim to reduce the confidence regions of $\hat{\theta}$.

Design criteria are mathematical functions, that depend on an experimental design. On the basis of these criteria one design can be judged better than another design. The reliability of the parameters depends on the size and the orientation of this ellipsoidal region. As a consequence, design criteria can be expressed as geometrical properties of the ellipsoidal region as will be shown in Section 5.4.

Another motivation for optimal experiment design is brought up in this chapter. If we want to discriminate between two models, which both fit the data, and we cannot discriminate on the basis of the available data, then information from the newly designed experiments should enable us to perform the discrimination between the given models.

5.3 Parameter - state variable dependence

In the majority of the parameter estimation problems not all unknown parameters can be estimated within acceptable bounds. Before we continue we should make the expression ‘acceptable’ more precise. From a naive point of view one might come up with the idea that, after calculating the individual confidence regions of each parameter, these confidence regions should be smaller in size than some predefined value, given a certain confidence level. This is not a good approach and we will try to explain this in the next paragraph.

One of the main goals of parameter estimation is to obtain a reliable model to study the physical process under consideration by performing simulations. This means that we should focus on the state variables which are of interest for physical reasons and how they relate to the separate parameters. Parameters which do not have great influence on the simulation results of the state variables of interest, do not need tight confidence limits and vice versa. Whether a confidence region is acceptable depends on the points of interest of the modeller.

This section introduces an approach to investigate parameter-state variable dependences by deriving quantities which describe the influence of a change in the j -th param-

eter on the i -th state variable. Reasoning in the reverse direction leads to the proposition that these quantities also indicate to which extent measurements of the i -th state variable lead to more accurate estimators of the j -th parameter. In the reverse case this quantity is corrected by the weight which corresponds to the i -th state variable. As in Section 3.3.2, we will assume that the variances of the measurement errors –and therefore the corresponding weights in (1.4)– are equal if they correspond to the same component of the state vector, $y(t, \theta)$.

In order to represent the information on the interactions clearly, we construct a labelled, bipartite graph $G = (P, L)$, where P is a set of vertices and L a set of edges connecting the elements of P . The set of vertices can be divided into two disjunct sets, P_1 and P_2 , containing the n dependent state variables and m parameters, respectively. Consequently, the graph will have a maximum of mn edges. The edge (y_i, θ_j) is an element of the set L , if the corresponding dependence is non-zero.

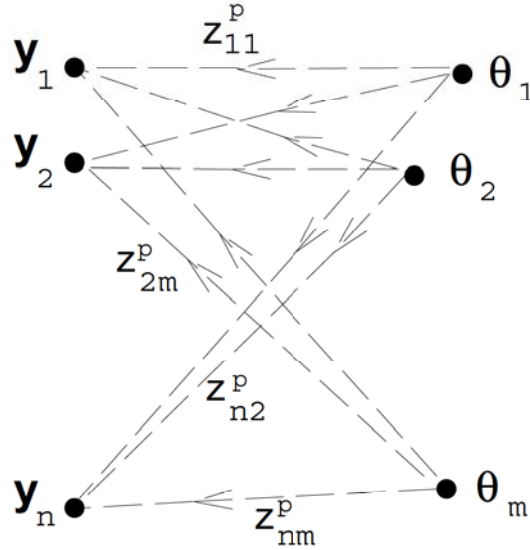


Figure 5.2: Graph to represent the dependences between state variables and parameters.

The labels, expressing the magnitude of the sensitivity of the j -th parameter on the i -th state variable, are defined as:

$$z_{ij}^{(p)} \stackrel{\text{def}}{=} \frac{\hat{\theta}_j}{\max_{t_0 \leq t \leq t_{\text{end}}} |y_i(t, \hat{\theta})|} \left\{ \frac{1}{t_{\text{end}} - t_0} \int_{t_0}^{t_{\text{end}}} \left| \frac{\partial y_i}{\partial \theta_j} \right|^p dt \right\}^{\frac{1}{p}}, \quad (5.1)$$

where the ratio $\hat{\theta}_j / \max_t |y_i(t, \hat{\theta})|$ is added to make the dependencies scale invariant, and $[t_0, t_{\text{end}}]$ is the time interval of the experiment. The derivatives $\partial y_i / \partial \theta_j$ are also called sensitivities. The pairwise dependences can be measured in many ways, we take the L^p -norm, with $1 \leq p \leq \infty$.

Apart from the quantification of the sensitivity of the i -th state variable on the j -th parameter, the labels as defined in (5.1) also have a reverse interpretation if they are corrected with the corresponding deviations. The correction reads:

$$\tilde{z}_{ij}^{(p)} \stackrel{\text{def}}{=} \frac{1}{\sigma_i} z_{ij}^{(p)}. \quad (5.2)$$

The corrected label, $\tilde{z}_{ij}^{(p)}$, indicates the influence of measurements of the i -th state variable on the j -th parameter.

The entry $\tilde{z}_{ij}^{(p)}$ can be seen as a scale invariant average over $[t_0, t_{\text{end}}]$ of all possible entries which might show up in the j -th column of the Jacobian (cf. (1.11)) after performing a measurement of the i -th component.

Remark 5.3.1 If the matrix $Z^{(p)}$ has a row whose elements are all zero, then the corresponding state variable is not dependent on any of the parameters. Measurements of these components will not contribute to more reliable estimates of the parameters.

Remark 5.3.2 If the matrix $Z^{(p)}$ has a column whose elements are all zero, then the corresponding parameter will have no influence on the model responses and can therefore not be estimated.

Example

In the case of the Barnes' problem (cf. Section 1.9 and Appendix 1.B), we have:

$$\begin{aligned} \hat{\theta} &= (0.861, 2.079, 1.815)^T, \\ t &\in [0, 6], \\ \max_t y_1(t, \hat{\theta}) &= 1.112, \\ \max_t y_2(t, \hat{\theta}) &= 0.585. \end{aligned}$$

After computing and integrating the sensitivities, we construct the matrix $Z^{(p)}$. The result for $p = 2$ reads:

$$Z^{(2)} = \begin{pmatrix} 0.394 & 1.015 & 0.879 \\ 0.835 & 1.118 & 0.575 \end{pmatrix}, \quad (5.3)$$

where we see that the biggest entries are in the second column, i.e. related to $\theta_2 = k_2$. The estimates are calculated with equal weights and 10 measurements of each component. When we consider the SVD (cf. (1.15)) of the corresponding Jacobian, the first

column of V equals $(-0.371, 0.746, -0.553)^T$, which is in agreement with the above results. From the matrix $Z^{(2)}$, we also see that y_1 is more sensitive to changes in k_3 and y_2 is more sensitive to changes in k_1 . Both conclusions are a bit surprising if we look into the equations and see that y'_1 and y'_2 only depend indirectly on k_3 and k_1 , respectively. \diamond

5.4 OED and improved confidence regions

In this section the target of optimal experiment design is to plan future experiments in such a way that the reliability of the parameter estimators, determined on the basis of previous and future experiments, will be optimal in some, mathematically well-defined sense. In order to study the reliability of the estimators we investigate the Jacobian of the regression problem, the design criteria depend on this matrix. To determine J in the case of linear regression we do not need a good estimate of θ . This is contrary to the nonlinear case, where we will need a good estimate for θ , in order to make a useful linearisation.

We assume that N measurements are already available and the corresponding least squares estimate is denoted by $\hat{\theta}$. (For optimal experiment design N may equal zero. In that case, $\hat{\theta}$ is an initial guess.)

Besides the N known measurements, we assume that a finite number of additional measurements, N_{add} , will be performed in the future. The final $(N + N_{add}) \times m$ Jacobian is denoted by

$$\check{J} = \left(\frac{J}{J_{add}} \right), \quad (5.4)$$

and $\check{\lambda}_1, \dots, \check{\lambda}_m$ are its positive, singular values in non-increasing order as in (1.15).

5.4.1 Design criteria

For different values of $\kappa \in [-\infty, +\infty]$, different design criteria can be distinguished, which are denoted by $\Psi_\kappa(\check{J}^T \check{J})$. If \check{J} has full rank:

$$\Psi_\kappa(\check{J}^T \check{J}) := \begin{cases} \check{\lambda}_1, & \kappa = +\infty, \\ \left(\frac{1}{m} \text{Tr}((\check{J}^T \check{J})^\kappa) \right)^{\frac{1}{\kappa}}, & \kappa \notin \{-\infty, 0, +\infty\}, \\ (\text{Det}(\check{J}^T \check{J}))^{\frac{1}{m}}, & \kappa = 0, \\ \check{\lambda}_m, & \kappa = -\infty, \end{cases} \quad (5.5)$$

and in the case \check{J} is singular:

$$\Psi_\kappa(\check{J}^T \check{J}) := \begin{cases} \check{\lambda}_1, & \kappa = +\infty, \\ \left(\frac{1}{m} \text{Tr}((\check{J}^T \check{J})^\kappa) \right)^{\frac{1}{\kappa}}, & \kappa \in]0, +\infty[, \\ 0, & \kappa \in [-\infty, 0]. \end{cases} \quad (5.6)$$

The determinant and trace of a matrix are abbreviated by Det and Tr, respectively. The design criterion, Ψ_κ , has special names for certain values of κ :

- D-optimal ($\kappa = 0$).
Here we maximise the determinant of the matrix $\check{J}^T \check{J}$, which is equivalent to minimising the volume of the ellipsoidal confidence region (4.11). A disadvantage of this choice of κ is the chance of constructing ‘thin and elongated’ confidence regions.
- A-optimal ($\kappa = -1$).
This choice of κ is equivalent to minimising the variance of $\sum_{i=1}^m \hat{\theta}_i$.
- E-optimal ($\kappa = -\infty$).
In this case we maximise the smallest singular value, $\check{\lambda}_m$, which means that we want to construct the ellipsoidal region in the parameter space as ‘sphere-shaped’ as possible.

In the case we are only interested in a subset of the parameters, because these parameters influence the simulation results of the state variables of interest, then we pre-multiply the Jacobian with a $m_A \times m$ -matrix ($m_A < m$) in order to zoom in on the more important parameters. The corresponding designs are known as D_A-, A_A- and E_A-design, the extensions to these designs are straightforward. More details with respect to design criteria may be found in [Loh93, Sil80].

Now, the final optimisation problem is to maximise $\Psi_\kappa(\check{J}^T \check{J})$ over N_{add} additional measurements, with N_{add} fixed. So we have to determine:

$$\max_{c_i, t_i (i=N+1, \dots, N+N_{add})} \Psi_\kappa(\check{J}^T \check{J}) , \quad (5.7)$$

and possible additional restrictions, which express experimental limitations:

$$\left. \begin{array}{l} g_1(c_i, t_i, N_{add}) = 0 \\ g_2(c_i, t_i, N_{add}) \leq 0 \end{array} \right\} \quad (i = N + 1, \dots, N + N_{add}) .$$

The maximum exists due to the facts that $t \in [t_0, t_{end}]$ and c_i and N_{add} are finite. In the next section we will show how to deal with the maximisation of the criterion function.

5.4.2 Repeated design

We assume that N is greater than zero and that every additional measurement has an experimental design such that for each $j = N + 1, \dots, N + N_{add}$, there is at least one $i = 1, \dots, N$, which meets: $\{c_i, t_i\} = \{c_j, t_j\}$. After N_{add} additional measurements have been performed, ω_i measurements under the i -th ($i = 1, \dots, N$) experimental design are available:

$$\sum_{i=1}^N \omega_i = N + N_{add} \quad \text{and} \quad \omega_i \geq 1 . \quad (5.8)$$

After the introduction of the diagonal matrix Ω , such that $(\Omega)_{ii} = \omega_i$, the overall Jacobian can be written as:

$$\check{J} = \Omega^{\frac{1}{2}} J . \quad (5.9)$$

- In the case of a repeated D-design we have to optimise:

$$\max_{\omega_1, \dots, \omega_N} \det(\check{J}^T \check{J}) = \max_{\omega_1, \dots, \omega_N} \det(J^T W J) = \det(J^T J) \max_{\omega_1, \dots, \omega_N} \prod_{i=1}^N \omega_i .$$

When we take the restrictions of (5.8) into account the maximum is attained if $\omega_i = 1 + N_{add}/N$. Because ω_i is an integer and, in general, $1 + N_{add}/N$ is not, some of the ω_i 's have to be rounded off in such a way that (5.8) is still fulfilled.

- In the case of a repeated A-optimal design we have to compute

$$\begin{aligned} \max_{\omega_1, \dots, \omega_N} \text{Tr}(\check{J}^T \check{J}) &= \max_{\omega_1, \dots, \omega_N} \text{Tr}(J^T W J) \\ &= \max_{\omega_1, \dots, \omega_N} \sum_{i=1}^N \omega_i \sum_{j=1}^m (J_{ij})^2 , \end{aligned}$$

which is a linear, integer programming problem. Adding the restrictions of (5.8) leads to the following strategy. Determine i^* such that $\sum_{j=1}^m (J_{i^*j})^2$ is maximal and for the frequencies we get

$$\omega_i = \begin{cases} N_{add} + 1 & \text{if: } i = i^* , \\ 1 & \text{otherwise.} \end{cases}$$

If there is not a unique i^* , any integer combination of the i^* 's will do.

- In the case of a repeated E-optimal design it is not possible to find a useful relation between the choice of ω_i and $\check{\lambda}_m$, because the SVD of \check{J} can be completely different from the SVD of J . A good solution is to determine the optimal repeated design by a sequential design as will be explained in the next section.

Definition 5.4.1 By an improved E-design we mean that $\check{\lambda}_{m-q} > \lambda_{m-q}$, where $q \in \{0, 1, \dots, m-1\}$ is the largest integer such that $\lambda_{m-q} = \lambda_{m-q+1} = \dots = \lambda_m$.

Theorem 5.4.1 If no improvement of the repeated E-design can be made: $\check{\lambda}_{m-q} = \lambda_{m-q}$, then $\check{\lambda}_{m-q} = \lambda_{m-q} = \dots = \lambda_m = 0$.

Proof: If a repeated design leads to an improved E-design, the repeated design with $\omega_i = 2$ ($i = 1, \dots, N$) leads to an improvement. The Jacobian after adding this design reads:

$$\check{J} = \begin{pmatrix} \omega_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \omega_N \end{pmatrix} J = 2J .$$

The singular value decomposition of this Jacobian equals:

$$\check{J} = U \check{\Sigma}^2 V^T ,$$

with $\check{\Sigma}^2 = 2\Sigma^2$ and U , Σ and V^T come from the SVD of the original Jacobian, J . When there is no improvement, it means that $\lambda_{m-q} = \check{\lambda}_{m-q}$ and by using the SVD of \check{J} we also have $\check{\lambda}_{m-q} = 2\lambda_{m-q}$. \square

5.4.3 Sequential design

The optimisation problem (5.7) is not solved directly, but we solve a slightly different problem. For this purpose, we take N_{add} equal to one, solve the minimisation problem and repeat this until some stopping criterion is fulfilled. Such approach is called sequential design [Fed72, page 173].

Sequential designs are much more attractive from a computational point of view, while asymptotically, $N_{add} \rightarrow \infty$, optimal sequential designs approach optimal nonsequential designs [Fed72]. The consequences for the design criteria as introduced in Section 5.4.1 in the case of sequential design are listed below, where J_{add} is a $1 \times m$ -matrix.

- In the case of sequential D-design we have to maximise:

$$\begin{aligned} \text{Det}(\check{J}^T \check{J}) &= \text{Det}(J^T J + J_{add}^T J_{add}) \\ &= \text{Det}(J^T J)(1 + J_{add}(J^T J)^{-1} J_{add}^T) , \end{aligned} \quad (5.10)$$

as a function of t and c_{N+1} . Maximising this determinant, by making use of the SVD of J leads to:

$$\max_{t_0 \leq t \leq t_{\text{end}}, c_{N+1}} \|J_{add} V \Sigma^{-1}\| . \quad (5.11)$$

- Sequential A-design leads to maximising:

$$\begin{aligned} \text{Tr}(\check{J}^T \check{J}) &= \text{Tr}(J^T J + J_{add}^T J_{add}) \\ &= \text{Tr}(J^T J) + \text{Tr}(J_{add}^T J_{add}) \end{aligned} \quad (5.12)$$

again as a function of the design variables t and c_{N+1} . The maximum of this sum of traces is attained at the same point as:

$$\max_{t \leq t \leq t_{\text{end}}, c_{N+1}} \|J_{add}\| . \quad (5.13)$$

- For the sequential E-design, where we want to improve the design by a max-min criterion on the singular values of the Jacobian, we have the following results.

Theorem 5.4.2 In the case of E-design, the criterion function is, after adding one additional measurement, bounded by

$$\lambda_{m-q-1} \geq \Psi_{-\infty}(\check{J}^T \check{J}) = \check{\lambda}_{m-q} \geq \lambda_{m-q} , \quad (5.14)$$

where q is taken as in Definition 5.4.1.

Proof: The additional row can be expressed in the columns of V :

$$J_{add}^T = \sum_{i=1}^m \beta_i V_i .$$

The matrix $\check{J}^T \check{J}$ can then be written as:

$$\check{J}^T \check{J} = V(\Sigma^2 + B)V^T , \quad (5.15)$$

where the i, j -th entry of the $m \times m$ -matrix B reads $\beta_i \beta_j$. Because the matrix V is orthogonal, the eigenvalues of $\check{J}^T \check{J}$ are the same as those of $\Sigma^2 + B$. Further, the matrix Σ^2 is diagonal and B has rank 1. Now the proof is easily completed by making use of the pages 433-434 of Golub and Van Loan, [GV83]. \square

Remark 5.4.1 A consequence of Theorem 5.4.2 is that the number of singular values of the Jacobian one wants to increase is equal to the minimal number of additional measurements to be performed in order to achieve this.

Remark 5.4.2 If $\beta_{m-q} = \dots = \beta_m = 0$, then there is no improvement of the E-design. In the next theorem we show that the reverse is also true.

Theorem 5.4.3 If no improvement for the sequential E-design can be constructed then $\beta_{m-q} = \dots = \beta_m = 0$.

Proof: The first part of the proof deals with the restrictive case where $q = 0$, i.e. $\lambda_{m-1} > \lambda_m$, and is proved by contradiction. Therefore, $\check{\lambda}_m = \lambda_m$ and we assume that $\beta_m \neq 0$. In the second part we deal with the case where $q > 0$.

If λ_m^2 is an eigenvalue of $\check{J}^T \check{J}$, it is also an eigenvalue of $\Sigma^2 + B$ (cf. (5.15)). This means that:

$$\begin{pmatrix} \beta_1^2 + \lambda_1^2 - \lambda_m^2 & \dots & \beta_1 \beta_{m-1} & \beta_1 \beta_m \\ \vdots & \ddots & \vdots & \vdots \\ \beta_1 \beta_{m-1} & \dots & \beta_{m-1}^2 + \lambda_{m-1}^2 - \lambda_m^2 & \beta_{m-1} \beta_m \\ \beta_1 \beta_m & \dots & \beta_{m-1} \beta_m & \beta_m^2 \end{pmatrix}$$

is singular. By the assumption $\beta_m \neq 0$, we can take the i -th row and subtract β_i/β_m ($i = 1, \dots, m-1$) times the m -th row. The determinant of the resulting matrix equals $\beta_m^2 \prod_{i=1}^{m-1} (\lambda_i^2 - \lambda_m^2)$ and should be zero, which completes the contradiction.

For the case $q > 0$ the contradiction is constructed by assuming that at least one $\beta_i \neq 0$ ($i = m-q, \dots, m$) and that λ_{m-q}^2 should have an algebraic multiplicity of $q+1$ in the characteristic polynomial of $\Sigma^2 + B$. \square

Remark 5.4.3 In the case no improvement of the sequential E-design can be constructed, then any nonsequential design will fail.

Theorem 5.4.4 If $\beta_{m-q} = \dots = \beta_m = 0$, then $\lambda_{m-q} = \dots = \lambda_m = 0$.

Proof: By Theorem 5.4.3 we know that $\beta_{m-q} = \dots = \beta_m = 0$ implies that no improvement of the E-design exists. By assuming that $\lambda_{m-q} = \dots = \lambda_m > 0$ we get the contradiction by using Theorem 5.4.1 and stating that then a repeated design with $\omega_i = 2$ ($i = 1, \dots, N$) would have given an improvement of the E-design. \square

Remark 5.4.4 Intuitively one might think that a design which leads to a maximal $|\beta_m|$ is an optimal sequential E-design. This is not true, which can be demonstrated by a simple counter example. Suppose that the Jacobian reads:

$$J = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},$$

and we can either take a sequential design with: $J_{add}^{(1)} = (10, 10)$ or $J_{add}^{(2)} = (0, 2)$. For the first design we get: $\lambda_1^{(1)} = 14.23$ and $\lambda_2^{(1)} = 1.58$, and for the second design: $\lambda_1^{(2)} = \sqrt{5} = 2.24$ and $\lambda_2^{(2)} = 2$.

Now we can only state that for an optimal sequential E-design $\beta_i \neq 0$ for at least one $i = m-q, \dots, m$, but we did not manage to find a sufficiently simple relation between J , or its SVD, and J_{add} . As far as we can see we need a SVD of $\Sigma^2 + B$ for every candidate of J_{add} , which is an infeasible approach.

From a computational point of view the sequential A-design is very attractive, because –contrary to sequential D- and sequential E-optimal design– an update of the singular value decomposition is not needed after adding a measurement. Sequential A-design is related to a workable expression, (5.11), and is therefore more attractive than sequential E-design. In practice the optimisation can be performed by a program for Lipschitzian global optimisation such as one whose implementation is described in [Pin95]. When the model equations are given by a set of DAEs, we choose a regular grid in time, solve the model and sensitivity equations, and store the corresponding solutions for each grid point. This approach significantly reduces the computation time of the DAE solver during the optimisation.

5.5 OED and model discrimination

As in Appendix 1.C we have two models, $y(t, \theta)$ and $z(t, \phi)$, and their corresponding estimates $\hat{\theta}$ and $\hat{\phi}$, respectively. We order the vectors y and z , such that their first k entries, correspond to the common, observable state variables. If we cannot discriminate between two models on the basis of an F-ratio test from Appendix 1.C, then we want to perform additional measurements in order to obtain a decisive result. In the case of a sequential design it is a straightforward way to compute:

$$\max_{i=\{1, \dots, k\}, t \in [t_0, t_{\text{end}}]} |y_i(t, \hat{\theta}) - z_i(t, \hat{\phi})|.$$

For a design where this absolute difference is maximal, it is *not* expected that the change in $S(\hat{\theta}) - S(\hat{\phi})$ is maximal after adding the corresponding measurement. The absolute difference should be corrected with the variances of $y_i(t, \hat{\theta})$ and $z_i(t, \hat{\phi})$ in such a way that it is unlikely for the additional measurement to end up right between the two model responses. The derivation of the variance of $y_i(t, \hat{\theta})$ after a measurement has been added is given by:

$$\begin{aligned} \text{var}(y_i(t, \hat{\theta})) &= \mathbf{E} \left(y_i(t, \hat{\theta}) - \mathbf{E}(y_i(t, \hat{\theta})) \right)^2 \\ &\approx \mathbf{E} \left(\sum_{j=1}^m \frac{\partial y_i(t, \hat{\theta})}{\partial \theta_j} (\theta_j - \hat{\theta}_j) \right)^2 \\ &= \sum_{j=1}^m \sum_{l=1}^m \frac{\partial y_i(t, \hat{\theta})}{\partial \theta_j} \frac{\partial y_i(t, \hat{\theta})}{\partial \theta_l} \mathbf{E} \left((\theta_j - \hat{\theta}_j)(\theta_l - \hat{\theta}_l) \right) \\ &= \sigma^2 \frac{\partial y_i(t, \hat{\theta})}{\partial \theta} (J^T J)^{-1} \left(\frac{\partial y_i(t, \hat{\theta})}{\partial \theta} \right)^T. \end{aligned}$$

The inverse of $\tilde{J}^T \tilde{J}$ can be computed easily, because the SVD of J is available and we may use the relation (recall that $B = \beta \beta^T$ as in (5.15)):

$$(\tilde{J}^T \tilde{J})^{-1} = V (\Sigma^2 + B)^{-1} V^T = V \left(\Sigma^{-2} - \frac{\Sigma^{-2} \beta \beta^T \Sigma^{-2}}{1 + \beta^T \Sigma^{-2} \beta} \right) V^T.$$

The derivation of $\text{var}(z_i(t, \hat{\phi}))$ is identical. Thus, the criterion for model discrimination amounts to:

$$\max_{\substack{i = \{1, \dots, k\} \\ t \in [t_0, t_{\text{end}}]}} \begin{cases} y_i(t, \hat{\theta}) - \mu \sqrt{\text{var}(y_i(t, \hat{\theta}))} - z_i(t, \hat{\phi}) - \mu \sqrt{\text{var}(z_i(t, \hat{\phi}))} \\ \text{if: } y_i(t, \hat{\theta}) > z_i(t, \hat{\phi}), \\ z_i(t, \hat{\phi}) - \mu \sqrt{\text{var}(z_i(t, \hat{\phi}))} - y_i(t, \hat{\theta}) - \mu \sqrt{\text{var}(y_i(t, \hat{\theta}))} \\ \text{if: } y_i(t, \hat{\theta}) < z_i(t, \hat{\phi}), \end{cases} \quad (5.16)$$

where μ should be positive.

5.6 OED and nonlinearity

Here we only give an indication of experimental design for the reduction of the nonlinearity of the regression problem. This topic is much more difficult than the OED dealt with above and hardly touched in literature, but targeting for future research. Reduction of the nonlinearity through experimental design is only of interest if neither the planar assumption (cf. Section 4.4) holds, nor a reparametrisation of the model reduces the nonlinearity. If both requirements are met, we want to perform N_{add} additional measurements in such way that the resulting $\max(\Gamma^\perp, \Gamma^\parallel)$ is minimal.

In the case of a repeated design with $\omega_i = \omega$ ($i = 1, \dots, N$), both $\|\dot{\eta}_h\|$ and $\|\ddot{\eta}_h\|$ (cf. Section 4.4) will be a factor $\sqrt{\omega}$ larger and due to (4.27) and (4.28), both the normal curvature and the parameter-effect curvature become a factor $\sqrt{\omega}$ smaller.

A more thorough investigation would be desirable, but goes beyond the reach of this thesis. We will end this section by a simple example where we compute two designs, one for the reduction of the nonlinearity and one for an increase of the reliability of the parameters. It turns out that these two designs are incompatible.

Example

We return to the example of Section 4.4, $y(t, \theta) = \exp(-\theta t)$. We have performed already two measurements at $t_1 = 1$ and $t_2 = 2$, and want to perform one additional measurement at t_3 . The Jacobian, with this additional measurement, reads:

$$\tilde{J} = \left(-\exp(-\hat{\theta}), -2\exp(-2\hat{\theta}), -t_3 \exp(-t_3 \hat{\theta}) \right)^T.$$

Because of the size of $\tilde{J}^T \tilde{J}$, the A-, D- and E-optimal design coincide and equal $t_3 = 1/\hat{\theta}$. Computation of the nonlinearity measures and minimising them leads to $t_3 = 0$, which is not a surprise if we look at the model equations. Except that this choice reduces the nonlinearity, it does not give any additional information related to the estimate. \diamond

5.7 Concluding remarks

In this chapter we give an outline of optimal experiment design. The topic of OED is relevant when the parameters are estimated, but some questions with respect to the model are not sufficiently resolved. Answers to these questions are relevant to improve the accuracy of model simulations, to discriminate between different models or to reduce the nonlinearity of the regression problem.

We introduced a method to quantify the dependencies between parameters and state variables, and to represent them in a clear way. Then it is shown that these dependencies are also of interest for the design of future experiments. If we want to improve the reliability of the parameter estimators, we have different mathematical criteria to determine whether an experimental design is optimal in a well defined sense. Depending on the

criterion, we derived the related D-, A- or E-optimal design for a repeated and for a sequential design. For the so-called E-design, it turned out to be difficult to determine the corresponding optimal design, although we managed to derive a number of results which are of practical interest in this context.

Experimental design in order to discriminate between models is also considered. For this aspect not only the maximal absolute differences between the model responses are of interest, but also the corresponding variances. A relation between experimental design and nonlinearity of the regression problem is also given in this chapter. However, here still many open questions for research exist. By means of an example we showed that different design criteria may give rise to incompatible designs.

Chapter 6

Case Studies

In this chapter we apply the techniques from Chapters 1 to 5 to solve a number of real-life problems which originate from a wide range of application areas. The problems were solved in close cooperation with scientists working in these application areas, because, for the evaluation of the many possible models, a good domain knowledge about the problem studied is indispensable. For a fruitful and efficient cooperation some of this knowledge is also required for the modeller, whereas, some mathematical background is needed for the scientist who is interested in a good mathematical model of the process he/she studies. Such multi-disciplinary cooperation requires a good interaction and it is our experience that efficient means of communication are prerequisite if the parties are working at geographically distant locations.

Each section in this chapter deals with a different problem. The problem in Section 6.1 was provided by an industrial partner and describes the formation of resins. Two examples from bio-chemistry on blood coagulation and plant cell growth are discussed in Sections 6.2 and 6.3, respectively. Section 6.4 describes a problem from Akzo Nobel research, where besides the parameter estimation problem also various steps of the modelling process are outlined. Another case study from the same research department is given in Section 6.5. It describes water penetration in an aramide yarn, which is modelled by a 1-dimensional PDE. Section 6.6 is devoted to an example from macroeconomic time series and compares the performance of existing ARMA and SETAR methods, with less general models which have fewer parameters. In the last section, 6.7, we solve a complex parameter estimation problem from chemical engineering, known from literature [BDB86], and compare our results with those from this paper.

6.1 Production of resins

6.1.1 Introduction

In this section we present a study on parameter estimation in the field of *resin* production. The model describes a mechanism of *methylolation* of melamine by formaldehyde. The methylolation is reversible, nine methylol melamines can be identified. Condensation is not considered. For details on this chemical process we refer to [GHW66].

The mathematical model of the chemical process contains a set of 12 differential

algebraic equations (DAEs) and 16 unknown parameters; 8 series of measurements are available, performed under different initial conditions and at different temperatures. To estimate the unknown parameters we apply the strategy as described in Chapter 1. With the available measured data, 12 of the 16 unknown parameters could be estimated within acceptable statistical bounds. In this study we show the effects of a reparametrisation of the model.

6.1.2 Reaction mechanism

A schematic representation of the chain of reactions of interest is given in Figure 6.1. In this figure we give a label, 'a'-'k', to each chemical component of interest; *formaldehyde* is represented by an 'o' and has no label. The meaning of the labels is given in Table 6.1.

label	symbol	full name
a	<i>melSol</i>	solid melamine
b	<i>melAq</i>	dissolved melamine
c	<i>mon</i>	mono-methylol melamine
d	<i>di</i>	N,N'-di-methylol melamine
e	<i>NN</i>	N,N-di-methylol melamine
f	<i>tri</i>	N,N',N''-tri-methylol melamine
g	<i>NNN'</i>	N,N,N'-tri-methylol melamine
h	<i>tet</i>	N,N,N',N''-tetra-methylol melamine
i	<i>NNN'N'</i>	N,N,N',N'-tetra-methylol melamine
j	<i>pen</i>	penta-methylol melamine
k	<i>hex</i>	hexa-methylol melamine

Table 6.1: Labels, symbols and full names of the chemical components.

Most reactions in the model involve the binding and loosening of formaldehyde. The reaction rates which correspond to the binding have a positive subscript. Negative subscripts indicate the reverse reaction rates. The subscript of a reaction rate is 2 when the binding of formaldehyde is next to another formaldehyde element and 1 otherwise (when the binding is on a free stick of λ , see Figure 6.1).

The reaction mechanism between *melamine* in its solid and dissolved form (labeled 'a' and 'b' respectively, in the figure) is unknown. This causes a less straightforward modelling of the process. The adaptations and assumptions we made to overcome this inconvenience are discussed Section 6.1.5.

For cyclic chemical reaction parts the product of the reaction rates corresponding to the clockwise part should equal the product of the reaction rates anti-clockwise. From the reaction scheme we see that this condition is fulfilled automatically.

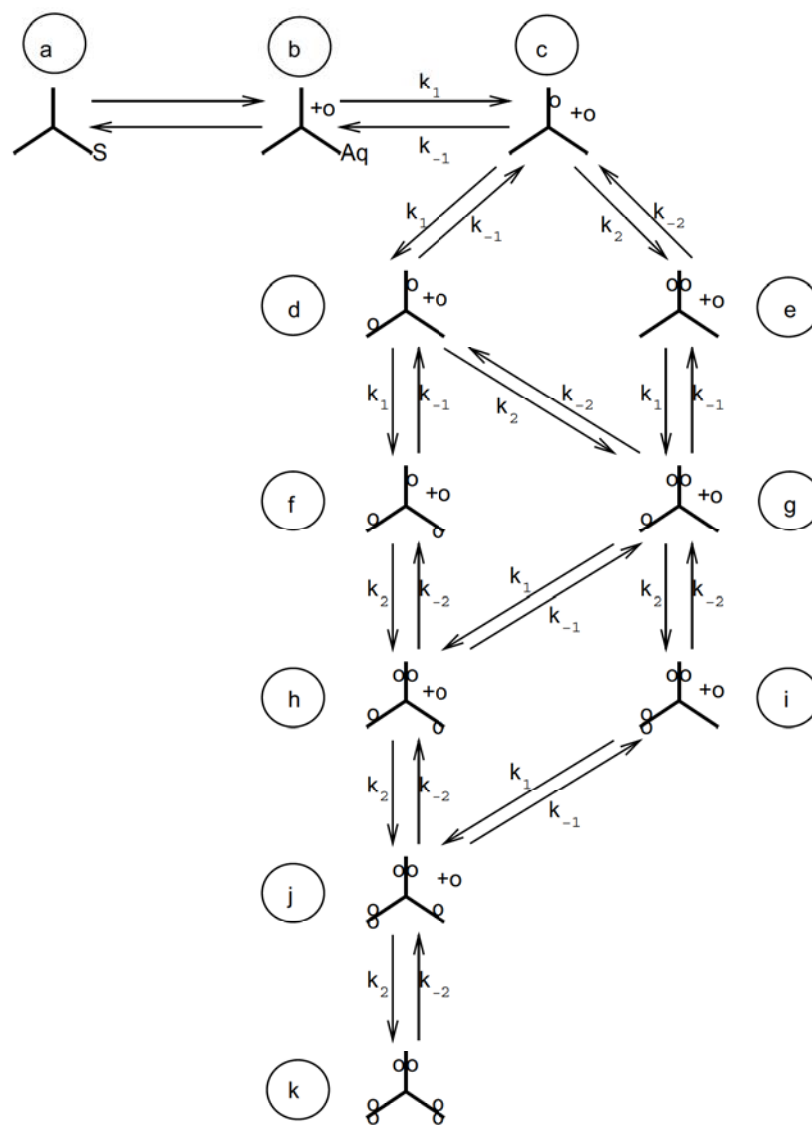


Figure 6.1: Scheme of the chain of reactions involved in the reversible methylation of melamine by formaldehyde. The labels 'a'-'k' are explained in Table 6.1.

6.1.3 Experiments performed

Eight series of measurements were performed under different initial conditions and at different temperatures. During each series, at a sequence of times, a sample of the reaction volume was taken, in which the formaldehyde concentration and the concentrations of the components with the labels 'b' to 'k' were measured. Each measurement gives the value of the concentration of one chemical component at a specific time, i.e. at each point of time we have 11 measurements. The total number of measurements (N) equals 583.

6.1.4 Model equations

Each differential equation in the mathematical model corresponds to a changing concentration of a chemical species. The derivation of the equations is based on straightforward second order reaction kinetics and on conservation of mass.

For illustration we focus on the formation, i.e. the change of concentration per unit of time, of mono-methylol melamine (label 'c') out of dissolved melamine (label 'b') and formaldehyde. This production depends on k_1 , on the concentrations of formaldehyde and dissolved melamine and on the number of possibilities for the binding of formaldehyde to dissolved melamine. In this case there are six places where the formaldehyde can be bound. The reverse reaction depends on k_{-1} , and on the concentration mono-methylol melamine and water. For this reverse step we only have one possibility for the loosening. Following these rules for the reaction kinetics and denoting the formaldehyde concentrations with $[FM]$, the water concentration with $[H_2O]$ and the concentration of a methylol melamine by its symbol (see Table 6.1) inside square brackets, we can derive the differential equations for all the species with the labels 'c' to 'k', as well as for formaldehyde and water. The resulting differential equations read:

$$\begin{aligned} \frac{d[FM]}{dt} = & -k_1[FM] (6[meAq] + 4[mon] + 2[di] + \\ & 4[NN] + 2[NNN] + 2[NNN'N']) - \\ & k_2[FM] ([mon] + 2[di] + 3[tri] + [NNN'] + 2[tet] + [pen]) + \\ & k_{-1}[H_2O] ([mon] + 2[di] + 3[tri] + [NNN'] + 2[tet] + [pen]) + \\ & k_{-2}[H_2O] (2[NN] + 2[NNN] + 2[tet] + \\ & 4[NNN'N'] + 4[pen] + 6[hex]) , \end{aligned} \quad (6.1)$$

$$\frac{d[H_2O]}{dt} = -\frac{d[FM]}{dt} , \quad (6.2)$$

$$\begin{aligned} \frac{d[mon]}{dt} = & 6k_1[FM][meAq] + 2k_{-1}[H_2O][di] + 2k_{-2}[H_2O][NN] - \\ & 4k_1[FM][mon] - k_2[FM][mon] - k_{-1}[H_2O][mon] , \end{aligned} \quad (6.3)$$

$$\begin{aligned} \frac{d[NN]}{dt} = & k_2[FM][mon] + k_{-1}[H_2O][NNN'] - \\ & 4k_1[FM][NN] - 2k_{-2}[H_2O][NN] , \end{aligned} \quad (6.4)$$

$$\begin{aligned} \frac{d[di]}{dt} = & 4k_1[FM][mon] + 3k_{-1}[H_2O][tri] + 2k_{-2}[H_2O][NNN'] - \\ & 2k_1[FM][di] - 2k_2[FM][di] - 2k_{-1}[H_2O][di] , \end{aligned} \quad (6.5)$$

$$\begin{aligned} \frac{d[NNN']}{dt} = & 4k_1[FM][NN] + 2k_2[FM][di] + 4k_{-2}[H_2O][NNN'N'] + \\ & 2k_{-1}[H_2O][tet] - k_2[FM][NNN'] - 2k_1[FM][NNN'] - \\ & 2k_{-2}[H_2O][NNN'] - k_{-1}[H_2O][NNN'] , \end{aligned} \quad (6.6)$$

$$\frac{d[tri]}{dt} = 2k_1[FM][di] + 2k_{-2}[H_2O][tet] - k_2[FM][tri] - 3k_{-1}[H_2O][tri] , \quad (6.7)$$

$$\begin{aligned} \frac{d[NNN'N']}{dt} = & k_2[FM][NNN'] + k_{-1}[H_2O][pen] - 2k_1[FM][NNN'N'] - \\ & 4k_{-2}[H_2O][NNN'N'] , \end{aligned} \quad (6.8)$$

$$\begin{aligned} \frac{d[tet]}{dt} = & 3k_2[FM][tri] + 2k_1[FM][NNN'] + 4k_{-2}[H_2O][pen] - \\ & 2k_2[FM][tet] - 2k_{-2}[H_2O][tet] - 2k_{-1}[H_2O][tet] , \end{aligned} \quad (6.9)$$

$$\begin{aligned} \frac{d[pen]}{dt} = & 2k_2[FM][tet] + 2k_1[FM][NNN'N'] - k_{-1}[H_2O][pen] - \\ & 4k_{-2}[H_2O][pen] + 6k_{-2}[H_2O][hex] - k_2[FM][pen] , \end{aligned} \quad (6.10)$$

$$\frac{d[hex]}{dt} = k_2[FM][pen] - 6k_{-2}[H_2O][hex] . \quad (6.11)$$

The concentrations are given in *mol/kg*, the time, *t*, in minutes and –hence– all reaction rates, *k_i*, in *kg/(mol min)*. These reaction rates, which are not known a priori, are the parameters to be estimated. We assume that the change of the reaction volume due to the dissolution of solid melamine may be neglected.

From the measurements we know that the temperature was not the same for all experiments. Therefore we account for a temperature dependence in the reaction rates by *Arrhenius' law*:

$$k_i(T) = \alpha_i \exp\left(\frac{-E_i}{RT}\right), \quad i \in \{-2, -1, 1, 2\}. \quad (6.12)$$

Here α_i is a *pre-exponential factor*, E_i the *activation energy*, R the *gas constant* and T the temperature (in Kelvin). By taking into account this temperature dependence, the number of unknown parameters is doubled.

To solve the set of differential equations (6.1)-(6.11), we need a set of corresponding initial conditions. These conditions describe the concentrations of the species of interest at the beginning of an experiment. We may assume that all initial concentrations are zero, except for water, formaldehyde and dissolved melamine (label 'b').

6.1.5 Treatment of the melamine concentrations

We already mentioned that the reaction mechanism between solid and dissolved melamine is unknown. This means that we are not able to derive an equation relating the concentrations of these species. On the other hand the concentration of dissolved melamine appears in the set of differential equations, which means that this concentration is indispensable for solving the differential equations. For each sample taken during the reaction

also the concentration of dissolved melamine has been determined. To obtain this concentration at the intervening time intervals we used a linear interpolation between the corresponding two subsequent measured concentrations of dissolved melamine.

This leads to a total of 11 differential equations, (6.1)-(6.11), and an algebraic equation due to the linear interpolation of the dissolved melamine concentration. The input file for the model equations, as it will be used by the spIds program [EHS95], is found in Appendix 6.A, at the end of this chapter.

6.1.6 Parameter estimation

The resulting system of differential algebraic equations (DAEs) contains eight unknown parameters (α_i and E_i) due to Arrhenius' law. For each series of experiments, besides these eight unknowns we also do not know the precise initial concentration of formaldehyde. Because we have eight series of measurements, we get eight extra unknown parameters: $[FM_i(t_0)]$, $i \in \{1, \dots, 8\}$.

For a convenient shorthand notation we introduce a 16-dimensional parameter vector θ and a 12-dimensional state vector, $y(t, \theta)$ of varying concentrations, depending on t and θ , as:

$$\theta = (\alpha_1, E_1, \alpha_{-1}, E_{-1}, \alpha_2, E_2, \alpha_{-2}, E_{-2}, [FM_1(t_0)], [FM_2(t_0)], [FM_3(t_0)], [FM_4(t_0)], [FM_5(t_0)], [FM_6(t_0)], [FM_7(t_0)], [FM_8(t_0)])^T, \quad (6.13)$$

$$y = ([melAq], [FM], [H_2O], [mon], [NN], [di], [NNN], [tri], [NNN^*N^*], [tet], [pen], [hex])^T. \quad (6.14)$$

The system of differential algebraic equations and the corresponding initial conditions are now denoted by:

$$E \frac{dy}{dt} = f(t, y, \theta), \quad y(t_0, \theta) = y_0(\theta), \quad (6.15)$$

where E is a diagonal, 12×12 matrix, with $(E)_{11} = 0$ and $(E)_{ii} = 1$ for $i \in \{2, \dots, 12\}$. This matrix E accounts for the distinction between differential and algebraic equations. The vector function $f : \mathbb{R} \times \mathbb{R}^{12} \times \mathbb{R}^{16} \rightarrow \mathbb{R}^{12}$ contains the information with respect to the linear interpolation (first component) and the differential equations for y_2, \dots, y_{12} (the right-hand sides of (6.1)-(6.11)). For details see Appendix 6.A.

6.1.7 Reparametrisation and results

The initial estimates for the pre-exponential factors and the activation energies (based on literature [GHW66]) and the initial formaldehyde concentrations (given by the experimentalists) are listed in Table 6.2. To obtain a better scaling of the numerical problem it is preferable to have the parameters within approximately the same order of magnitude. To achieve this we take the logarithm of the pre-exponential factors, α_i , and we scale

parameter	value	parameter	value
α_1	1.35×10^{14}	$FM_1(t_0)$	8.41
E_1	9.8×10^4	$FM_2(t_0)$	7.61
α_{-1}	3.98×10^8	$FM_3(t_0)$	5.60
E_{-1}	6.8×10^4	$FM_4(t_0)$	5.58
α_2	1.66×10^{15}	$FM_5(t_0)$	4.80
E_2	1.2×10^5	$FM_6(t_0)$	4.81
α_{-2}	8.91×10^9	$FM_7(t_0)$	4.80
E_{-2}	9.0×10^4	$FM_8(t_0)$	5.58

Table 6.2: Initial estimates for the unknown parameters.

the activation energies by a factor $1/1000$, $\tilde{E}_i = E_i/1000$. The scaled initial parameter estimates are listed in the second column of Table 6.3.

After the above scaling, the first numerical runs were performed by the approach described in Chapter 1. The results are reported in Table 6.3. A typical result is shown in Figure 6.2. The corresponding graphs of the calculated concentrations and the measured values of N,N',N''-tri-methylol melamine (label 'f' in Figure 6.1) during the second experiment and penta-methylol melamine (label 'j') during the eighth experiment for the initial and final parameter values are shown.

The results from Table 6.3, with respect to the sum of squares and the corresponding graphs are satisfactory; the numerical solution fits the measurements within reasonable bounds. However, the confidence regions for the pre-exponential factors and the activation energies are not satisfactory. The singular values and the columns of matrix V are shown in Figure 6.5. Inspection of the singular values (cf. Eq. (1.15)) shows that four of them are extremely small, see Figure 6.5. The corresponding singular vectors, the last four columns of V , can be identified with pairs $\{\ln(\alpha_i), \tilde{E}_i\}$, for $i \in \{-2, -1, 1, 2\}$. The same holds for the four largest singular values. This means that an intersection of the ellipsoidal region with the $\{\ln(\alpha_i), \tilde{E}_i\}$ -plane gives an elongated ellipse, of which the principal axes are rotated with respect to the coordinate axes. The presence of elongated ellipsoidal regions can also be seen from the ratios of the independent and dependent confidence regions. This indicates that for each pair $\{\ln(\alpha_i), \tilde{E}_i\}$, only one parameter can be estimated accurately after an appropriate reparametrisation of either $\ln(\alpha_i)$ or \tilde{E}_i . A plot of the intersection of the iso-curves of the sum of squared discrepancies with the $\{\ln(\alpha_1), \tilde{E}_1\}$ -plane is given in Figure 6.3. The elongated shapes in this figure are in accordance with what was expected after the linear investigation. Additional information comes from asymmetry in the north-west and south-east direction of this figure. This indicates the presence of nonlinear effects. In the remainder of this section we will show that this is due to parameter-effect curvature (cf. Section 4.4).

A well known *reparametrisation* for the pre-exponential factor (see [BDB86, Wat94])

	initial est. (θ_{ini})	final est. ($\hat{\theta}$)	independent confidence regions ($\Delta^I \theta$)	dependent confidence regions ($\Delta^D \theta$)
$\ln(\alpha_1)$	32.54	20.17	5.12	0.0728
\tilde{E}_1	98.00	65.38	14.0	0.198
$\ln(\alpha_{-1})$	19.80	24.81	20.5	0.469
\tilde{E}_{-1}	68.00	91.27	57.7	1.32
$\ln(\alpha_2)$	35.05	14.17	21.8	0.261
\tilde{E}_2	120.00	51.03	59.7	0.717
$\ln(\alpha_{-2})$	22.91	9.126	32.2	0.407
\tilde{E}_{-2}	90.00	47.61	88.4	1.13
FM_1	8.41	8.745	0.622	0.582
FM_2	7.61	8.536	0.609	0.578
FM_3	5.6	5.097	0.607	0.604
FM_4	5.58	6.098	0.712	0.701
FM_5	4.8	4.671	0.766	0.760
FM_6	4.81	4.724	0.768	0.752
FM_7	4.8	5.383	0.694	0.686
FM_8	5.58	6.065	0.702	0.683
$S(\theta)$	336.6	14.76		

Table 6.3: Initial estimates and final estimates of θ plus confidence regions (cf. (1.25) and (1.26) with $\alpha = 0.05$).

is found by introducing a *reference temperature*, T_0 . It leads to the formulation:

$$\begin{aligned}
 k_i(T) &= \alpha_i \exp\left(\frac{-E_i}{RT}\right) \\
 &= \tilde{\alpha}_i \exp\left(\frac{-E_i}{R} \left(\frac{1}{T} - \frac{1}{T_0}\right)\right), \quad i \in \{-2, -1, 1, 2\},
 \end{aligned} \tag{6.16}$$

with:

$$\tilde{\alpha}_i = \alpha_i \exp\left(\frac{-E_i}{RT_0}\right).$$

The temperature T_0 should be close to the temperatures during the experiments. An appropriate choice for T_0 is the average temperature over all the performed experiments. Note that the reparametrised pre-exponential factors, $\tilde{\alpha}_i$, represent the reaction rates, k_i , at $T = T_0$. The results after this reparametrisation are given in Table 6.4 for $T_0 = 333K$.

This reparametrisation does not change the model responses; the sum of squares and, except for α_i , the estimated parameter values are unaffected. Only the confidence

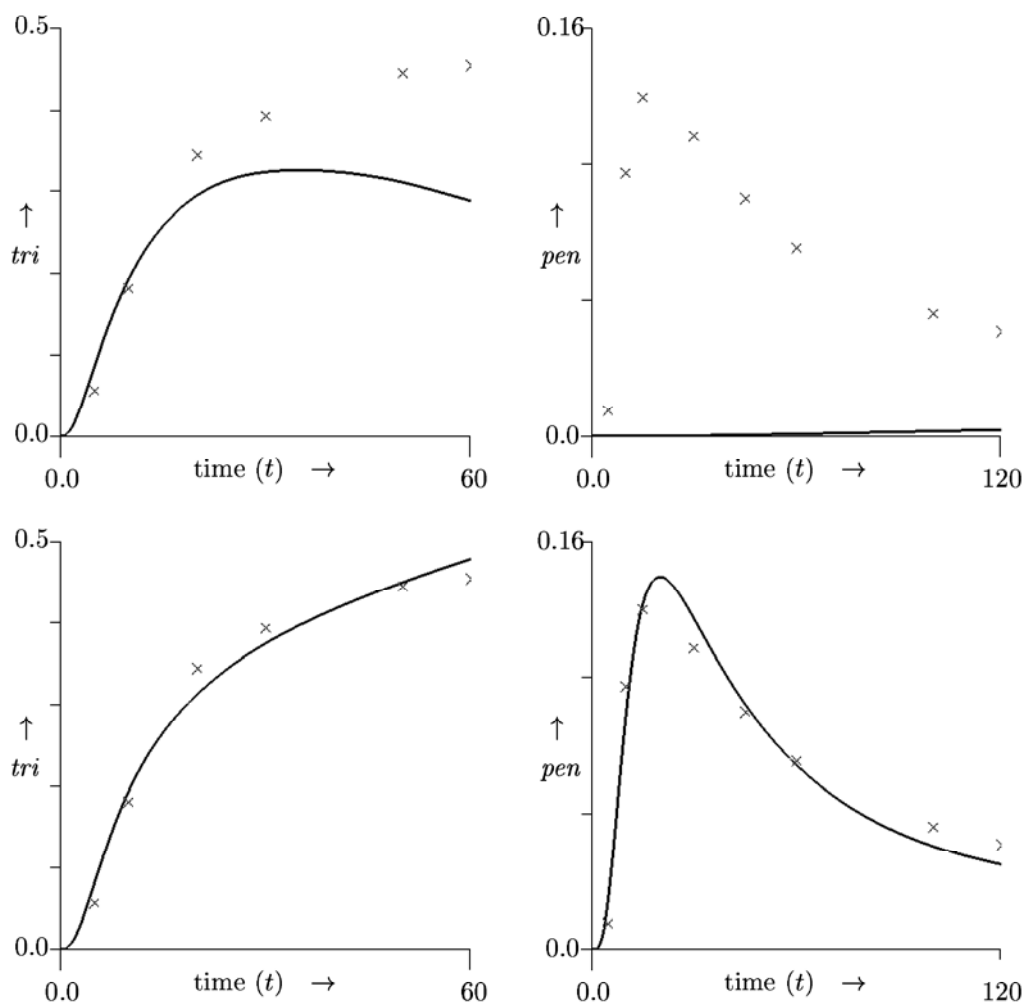


Figure 6.2: Measured ('x') and computed concentrations of N,N',N''-tri-methylol melamine (label 'f') during the second experiment (left) and the penta-methylol melamine (label 'j') during the eighth experiment (right), for the initial (top) and final (bottom) parameter values from Table 6.3.

regions of the reparametrised parameters improve. Inspection of the singular values shows again that four of them are extremely small. The essential difference with the results from Table 6.3 is that now the last four columns of the matrix V can be identified with the activation energies, E_i , i.e. the parameters which are the least well determined.

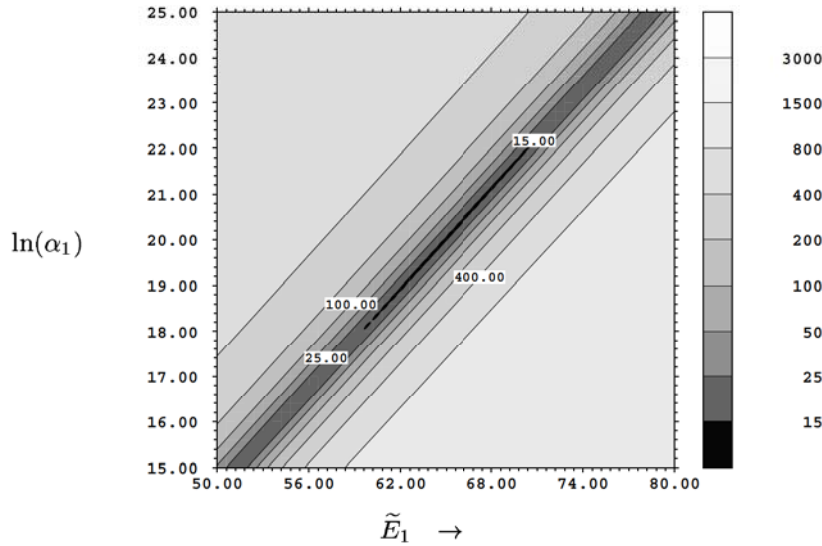


Figure 6.3: Level sets of the sum of squared discrepancies intersected with the $\{\ln(\alpha_1), \tilde{E}_1\}$ -plane, before the reparametrisation of (6.16).

This means that the longest principal axis of the elongated ellipse is rotated towards the E_i -axis by the reparametrisation. Level sets of the sum of squared discrepancies in the $\{\ln(\tilde{\alpha}_1), \tilde{E}_1\}$ -plane are shown in Figure 6.4. In this figure we see almost perfect ellipses which indicates that the problem is close to linear in its parameters after the reparametrisation of (6.16). According to the linear approximation (cf. (4.12)) with $\alpha = 0.05$ we get: $S(\theta) = S(\hat{\theta}) (1 + m/(N - m) \mathcal{F}_{0.05}(m, N - m)) = 15.47$. Comparison of the dependent confidence region of Table 6.4 and the intersections of the ellipse for $S(\theta) = 15.47$ in Figure 6.4 give a close correspondence; the distance from the centre of the ellipse to the intersections with the parameter axes are 0.074 and 7.21 for $\ln(\tilde{\alpha}_1)$ and \tilde{E}_1 , respectively.

The available measurements were carried out at temperatures between 323K and 353K. In order to estimate the parameters E_i more accurately, additional measurements are required which span a wider range of temperatures.

6.1.8 Conclusions

In this section we applied the parameter estimation approach as described in Chapter 1 to 5 to a real-life problem from reaction kinetics in order to estimate unknown

	initial estimates (θ_{ini})	final estimates ($\hat{\theta}$)	independent confidence regions ($\Delta^I \theta$)	dependent confidence regions ($\Delta^D \theta$)
$\ln(\tilde{\alpha}_1)$	-2.74	-3.376	0.134	0.073
\tilde{E}_1	98.00	65.33	14.0	7.38
$\ln(\tilde{\alpha}_{-1})$	-4.68	-8.047	0.65	0.467
\tilde{E}_{-1}	68.00	91.91	57.2	38.3
$\ln(\tilde{\alpha}_2)$	-8.15	-4.181	0.621	0.261
\tilde{E}_2	120.00	54.23	61.4	25.1
$\ln(\tilde{\alpha}_{-2})$	-9.49	-7.986	0.893	0.405
\tilde{E}_{-2}	90.00	53.03	88.9	38.2
FM_1	8.41	8.743	0.621	0.582
FM_2	7.61	8.534	0.608	0.578
FM_3	5.6	5.097	0.607	0.604
FM_4	5.58	6.097	0.712	0.702
FM_5	4.8	4.672	0.766	0.760
FM_6	4.81	4.723	0.768	0.752
FM_7	4.8	5.382	0.694	0.686
FM_8	5.58	6.065	0.703	0.683
$S(\theta)$	335.7	14.77		

Table 6.4: Initial and final estimates of θ , plus confidence regions (cf. (1.25) and (1.26) with $\alpha = 0.05$), after reparametrisation of the pre-exponential factor.

reaction rates and unknown initial concentrations. The experiments were performed at different temperatures, which made it necessary to use Arrhenius' law to describe the reactions rates. The unknown initial concentrations and pre-exponential factors could be determined, with an accuracy which was satisfactory to the experimentalists. For that purpose, however, we needed a reparametrisation of the pre-exponential factor. However, due to the small range of the temperatures for which experimental data were available, it was not possible to estimate the activation energies accurately.

The reparametrisation reduces the parameter-effect curvature and the intersection in Figure 6.4 is in full agreement with the results from linear statistics. Another advantage of the reparametrisation, which was encountered during the numerical experiments, is the decrease of the number of steps in the minimisation routine.

This example illustrates the strength of the method, which yields the capability to decide for which parameters sufficient information is available in order to perform an accurate estimation procedure. The visualisation as shown in Figure 6.5 turns out to be a convenient aid to see immediately the structure of the relevant information from the singular value decomposition.

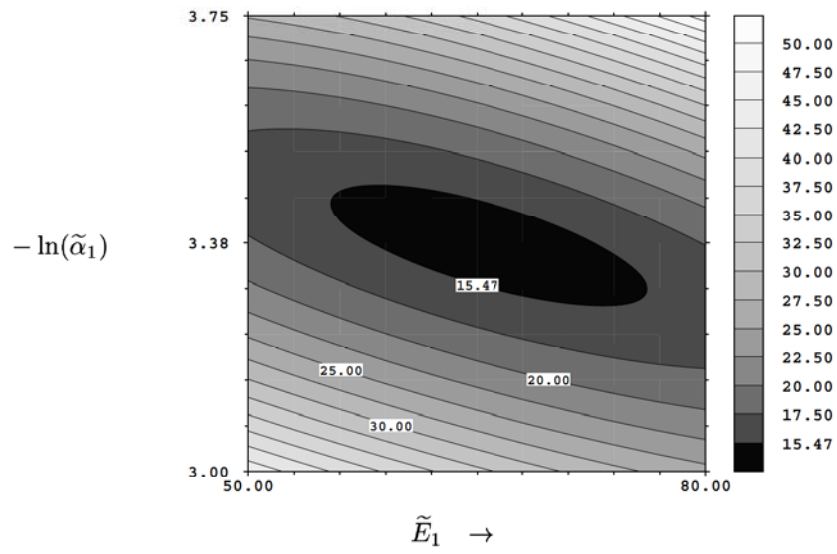


Figure 6.4: Level sets of the sum of squared discrepancies intersected with the $\{\ln(\tilde{\alpha}_1), \tilde{E}_1\}$ -plane, after the reparametrisation of (6.16).

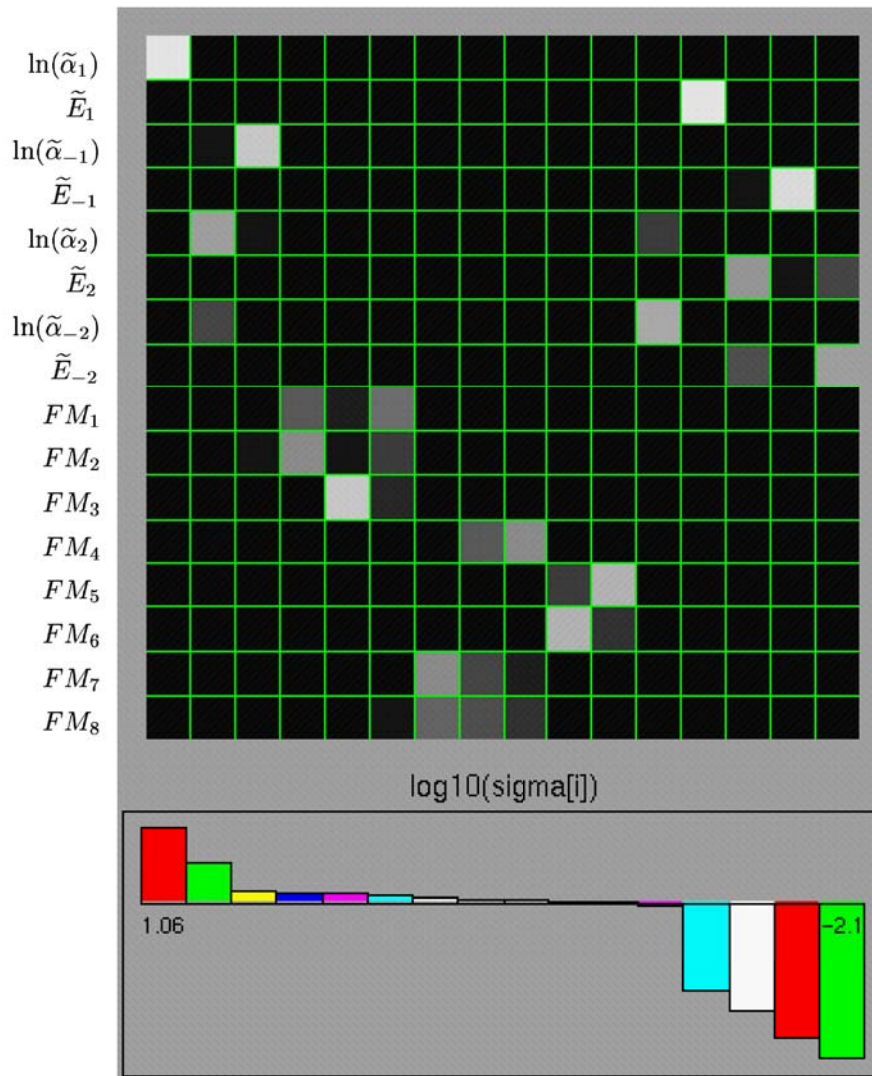


Figure 6.5: The squared entries of the 16×16 -matrix V are matched on a grey scale. The black squares indicate small values, the white squares represent values close to 1. The columns correspond with the singular values in decreasing order. The logarithms of the corresponding singular values are shown in a histogram at the lower part of the picture. The rows in the matrix correspond with the various parameters in the order given by (6.13).

6.2 Mathematical modelling in blood coagulation[†]

This section describes the mathematical modelling of a part of the blood coagulation mechanism. The model includes the activation of factor X by a purified enzyme from Russel's Viper Venom (RVV), factor V and prothrombin, and also comprises the inactivation of the products formed.

In this study we assume that in principle the mechanism of the process is known. However, the exact structure of the mechanism is unknown, and the process still can be described by different mathematical models. These models are put to test by measuring their capacity to explain the course of thrombin generation as observed in plasma after recalcification in presence of RVV. The mechanism studied is mathematically modelled as a system of differential-algebraic equations (DAEs). Each candidate model contains some freedom, which is expressed in the model equations by the presence of unknown parameters. For example, reaction constants or initial concentrations are unknown. The goal of parameter estimation is to determine these unknown parameters in such a way that the theoretical (i.e., computed) results fit the experimental data within measurement accuracy and to judge which modifications of the chemical reaction scheme allow the best fit.

We present results on model discrimination and estimation of reaction constants, which are hard to obtain in another way.

6.2.1 Introduction

One of the problems encountered in the study of a complicated biochemical process like thrombin generation in plasma, is that neither the reaction mechanism nor the reaction constants and initial concentrations are precisely known. The knowledge on the reaction mechanism of the process is obtained mainly through experiments on isolated parts of the system. The elements of the system, i.e. the clotting factors and their interactions, are separated from blood plasma and their interaction is studied under circumstances that are necessarily not precisely identical to those under which they cooperate in plasma. In fact it is not even known whether the reaction scheme that we deduce from such experiments is indeed the one operative in plasma. There may exist unknown factors or reactions, and reactions that have been shown to be possible in principle may not occur in reality. An example of this is the fact that factor X_a can activate factors V and VIII under experimental circumstances, but that this reaction does not seem to play a role in clotting plasma [MT90]. Also the reaction conditions in plasma are different from those used for the study of the interaction of isolated factors. They may even be unsuitable for the study of such interactions. The kinetic parameters of activation of factor V by thrombin, e.g., cannot be measured directly in plasma because the presence of natural thrombin inhibitors renders it impossible to achieve a fixed enzyme concentration.

[†] This section results from joint work with H.C. Hemker (Department of Biochemistry, University Maastricht) and P.W. Hemker (CWI, Amsterdam) and will be submitted in an almost identical form.

We introduce mathematical model validation and parameter estimation as a possible solution to these problems. In this procedure, on basis of the existing biochemical knowledge, a probable reaction mechanism is postulated. This is transformed into a set of differential-algebraic equations, which contains unknown parameters. These parameters correspond with the reaction constants and initial concentrations of the reactants, both approximately known from previous experiments and used as an initial guess for the parameters to be estimated. Then, one or more results of the reaction process are monitored, e.g. the course of thrombin concentration in plasma in time after triggering of the coagulation process, and the parameters in the model are adapted to obtain an optimal fit. Different hypothetical reaction mechanisms can be tested in parallel to see which one results in a better fit. If the best fit leads to improbably large discrepancies between the computed and the experimental results, the model is adapted and the validation process is repeated.

In this case study we briefly indicate this process of model derivation and validation. In fact, the process consists of checking a long sequence of improving models, adapted during the process for a wide range of reasons. The final model should not only lead to a satisfactory fit, but should also be simple, in accordance with established facts, and –preferably– it should not contain an unreasonably large number of parameters. In order to validate the many models and to estimate the corresponding parameters, an interactive software package for parameter estimation on a fast computer is an indispensable tool. Such a computer program, called *spIds* [EHS95] and partially constructed by two of the authors, was available to carry out the necessary computations.

The model we consider here describes thrombin formation, a part of the blood coagulation process, by a system of differential-algebraic equations. The variation in time of the concentrations of each reactant is described by a (differential) equation. The chain of reactions which leads to thrombin starts with the activation of factor X by RVV, followed by the activation of factor V, the production of prothrombinase in the presence of phospholipid and the activation of prothrombin. We also take into account the inactivation of the factor Xa by anti-thrombin III (ATIII) and the inactivation of thrombin by ATIII and α_2 -macroglobulin (α_2 M).

A description of the experiments used is given in Section 6.2.2, followed by a derivation of the reaction mechanism in Section 6.2.3. The step from reaction mechanism to mathematical equations is given in Section 6.2.4. The parameter estimation process is briefly described in Section 6.2.5. The results and conclusions are given in Sections 6.2.6 and 6.2.7, respectively.

6.2.2 Experimental data

In order to obtain the required data, four experiments were performed, which resulted in four series of measurements. The output of the system used for our tests was the course of thrombin-like amidolytic activity. This activity is caused by two types of molecules: thrombin itself and the thrombin- α_2 macroglobulin complex (briefly denoted as II_a and

$\text{II}_a - \alpha_2\text{M}$ respectively, in the reaction scheme, Figure 6.6).

The data were obtained as follows. To 240 μl of defibrinated plasma, in which the clotting factors are contained, we add 3.6 μl of a suspension of procoagulant phospholipids (1 μM) and 80.4 μl of a solution of RVV. This concentration of RVV was halved in the subsequent experiments. The thrombin generation process was started at $t = 0$ by addition of 36 μl of CaCl_2 (100 mM). At different time intervals, more frequently in the initial phase of the reaction and less frequently at the end, we took 0.01 ml samples from the reaction mixture and added it to 0.49 ml of a solution of the chromogenic substrate S2238 (0.5 mM) in a buffer that contains the Ca_2^+ chelating agent EDTA in order to stop further thrombin generation. Thrombin and $\alpha_2\text{M}$ -thrombin split the yellow-coloured para-nitroaniline from S2238. After 2 min. this reaction is stopped by adding citric acid and the colour is measured and used to determine the thrombin activity in the sample. Time measurements for the thrombin generation are made automatically and samples are taken until a stable end level of amidolytic activity is observed. This takes about 15 minutes.

6.2.3 Reaction mechanism

At this point we first present a commonly accepted reaction sequence for thrombin generation in Figure 6.6. Thereafter we describe three possible variants as found in [Hem93]. In this section the reaction mechanism and its alternatives are given in a schematic way. In Section 6.2.4 we give a more precise description by deriving differential equations. This is followed by an overview of the motivation and selection criteria involved in choosing one set of equations in favour of its alternatives.

In the reaction schemes the coagulation factors are denoted by their Roman numbers, the subscript ‘a’ indicates their activated form, ‘PL’ and ‘PT’ denote phospholipid and prothrombinase, respectively. ‘ATIII’ and ‘ $\alpha_2\text{M}$ ’ (anti-thrombin III and α_2 -macroglobulin) are responsible for inactivation of the factors II_a and X_a .

In the scheme of Figure 6.6, the activation of X by RVV, (reaction r_1), leads to X_a , followed by its inactivation by ATIII (r_2). Next, factor V is activated by II_a (r_3). The factors X_a , V_a and PL produce PT in a reversible association (r_4 and r_5). Subsequently, thrombin (II_a) is formed out of prothrombin (II), either in the presence of PT (r_6) or of X_a (r_7). Finally, II_a is inactivated either by $\alpha_2\text{M}$ or by ATIII (r_8 and r_9 , respectively).

In this study we show that the above scheme is suitable to explain the experimental results. It summarises the present common knowledge, but it is not necessarily complete and/or unique. We also investigate a number of possible alternatives. One such alternative concerns the formation of prothrombinase (PT), not in a trimolecular reaction but as a sequence of bimolecular reactions (Figure 6.7). Two other alternatives are given in the Figures 6.8 and 6.9. In the former we account for the existence of the intermediate meizothrombin that in itself has amidolytic activity [BTH⁺95], in the latter we account for the existence of an intermediate form of the $\alpha_2\text{M}$ -thrombin complex [MFG92]. All proposed alternatives are more complex than the reaction mechanism we start with in

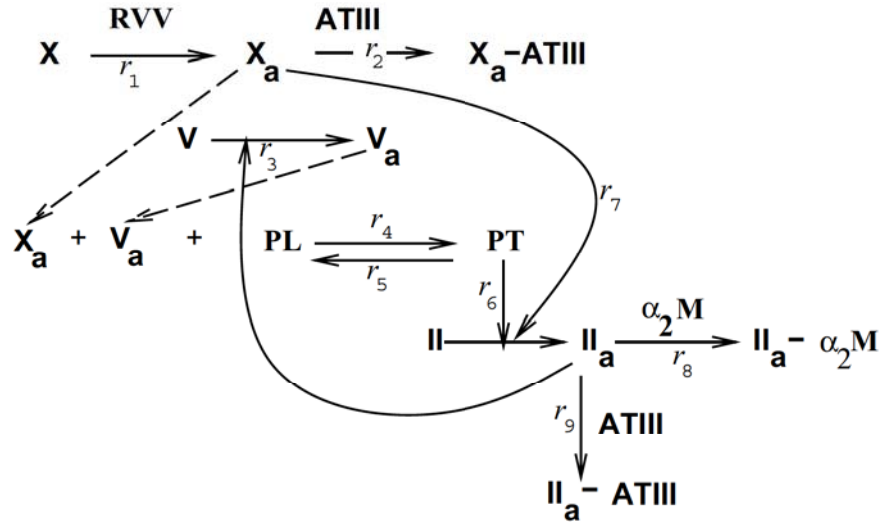


Figure 6.6: The reaction scheme for the part of the blood coagulation studied.

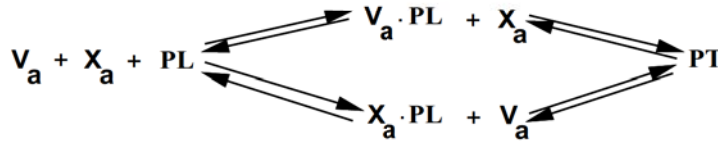


Figure 6.7: The alternative reaction scheme to account for prothrombin formation.

Figure 6.6. By ‘more complex’ we mean that it has more state variables and more intermediate reactions, which implies that they are likely to fit better because there are more degrees of freedom available. In Section 6.2.5 we will derive model equations from the reaction schemes and judge by statistical tests if an increase of the complexity of the model leads to a significant improvement of the fit between the calculated model responses and the observed data.

6.2.4 Model equations

From the four reaction schemes as they are introduced in Section 6.2.3, mathematical model equations were derived. It is obvious that the schemes presented lead to different

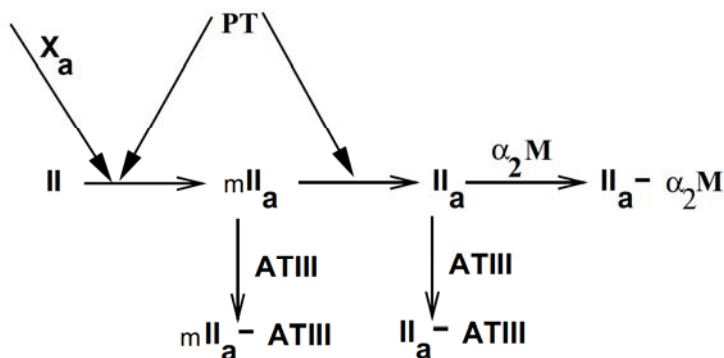


Figure 6.8: The alternative reaction mechanism for the formation of thrombin by the introduction of an intermediate reactant, meizothrombin (mII_a).

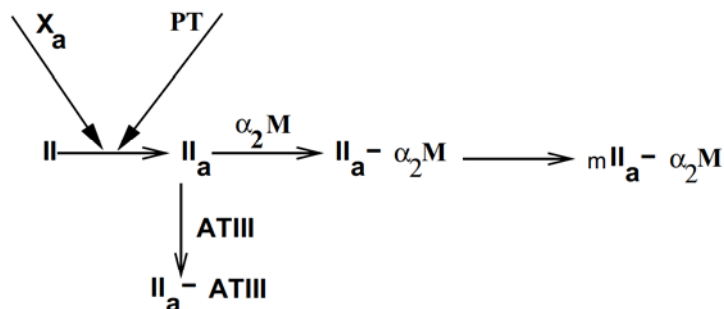


Figure 6.9: The alternative reaction mechanism for thrombin inactivation by $\alpha_2 M$. Here we assume that $II_a \alpha_2 M$ transforms further into an amidolytic less active form, $mII_a \alpha_2 M$.

sets of equations. But also from a single reaction scheme various sets of alternative mathematical model equations can be derived. As an example we consider the reaction r_1 , which is present in all four reaction schemes. The concentrations of the chemical species are given in nM and indicated by '[]'; the time, t , is given in minutes. The dimension of the reaction constants are derived from these units. The change in time of the concentration of factor X can be given by the well-known Michaelis-Menten relation:

$$\frac{d[X]}{dt} = -r_1 = -\frac{kcat_X \cdot [X] \cdot [RVV]}{km_X + [X]}. \quad (6.17)$$

Although we know from literature that this relation is likely to be valid, it may be replaced by closely related expressions. In cases where $km_X \gg [X]$ or $km_X \ll [X]$, expression (6.17) transforms respectively into the alternatives

$$r_1 = kk_1 \cdot [X] \cdot [RVV] , \quad (6.18)$$

with $kk_1 \approx kcat_X/km_X$ or

$$r_1 = kk_2 \cdot [RVV] , \quad (6.19)$$

with $kk_2 \approx kcat_X$. Both alternatives have one parameter less than the Michaelis-Menten relation and, depending on the ratio $km_X/[X]$, they can replace (6.17) without loss of accuracy. A third possible alternative reads:

$$r_1 = kk_3 \cdot [X] , \quad (6.20)$$

which follows from (6.18), when RVV -dependence is negligible. Similar alternatives exist for the other reactions. Together, this leads to a large number of candidate models.

From all these candidates we select that model (or subset of models, if the statistical tests do not lead to a decisive answer) which, (i) is in accordance with established knowledge in the field, (ii) is devoid of irrelevant steps (cf. the Michaelis-Menten reaction mentioned above), and (iii) fits the phenomena observed.

In Section 6.2.5 we will highlight the process of parameter estimation and deal with model validation. In the last part of the present section we give the set of model equations which was chosen from the candidates on the basis of the criteria (i)-(iii). This set is one of the possible mathematical representations for the scheme given in Figure 6.6. and as such it is an example of the many possible systems of DAEs. In addition, it describes the connection with the experiments.

The selected system of equations reads:

$$\frac{d[X]}{dt} = -r_1 , \quad (6.21)$$

$$\frac{d[Xa]}{dt} = r_1 - r_2 - r_4 + r_5 , \quad (6.22)$$

$$\frac{d[V]}{dt} = -r_3 , \quad (6.23)$$

$$\frac{d[Va]}{dt} = r_3 - r_4 + r_5 , \quad (6.24)$$

$$\frac{d[PL]}{dt} = -r_4 + r_5 , \quad (6.25)$$

$$\frac{d[PT]}{dt} = r_4 - r_5 , \quad (6.26)$$

$$\frac{d[II]}{dt} = -r_6 - r_7 , \quad (6.27)$$

$$\frac{d[IIa]}{dt} = r_6 + r_7 - r_8 - r_9, \quad (6.28)$$

$$\frac{d[IIa\alpha_2M]}{dt} = r_9, \quad (6.29)$$

$$AmAct = [IIa] + 0.556 \cdot [IIa\alpha_2M], \quad (6.30)$$

$$r_1 = \frac{kcat_X \cdot [X] \cdot [RVV]}{km_X + [X]}, \quad (6.31)$$

$$r_2 = ki_{Xa} \cdot [Xa], \quad (6.32)$$

$$r_3 = \frac{kcat_V \cdot [V] \cdot [IIa]}{km_V + [V]}, \quad (6.33)$$

$$r_4 = k_{PT} \cdot [Va] \cdot [Xa] \cdot [PL], \quad (6.34)$$

$$r_5 = k_{PL} \cdot [PT], \quad (6.35)$$

$$r_6 = \frac{kcat_{II} \cdot [II] \cdot [PT]}{km_{II} + [II]}, \quad (6.36)$$

$$r_7 = \frac{kcat_2 \cdot [II] \cdot [Xa]}{km_2 + [II]}, \quad (6.37)$$

$$r_8 = ki_{IIa\alpha_2M} \cdot [IIa], \quad (6.38)$$

$$r_9 = ki_{IIaATIII} \cdot [IIa]. \quad (6.39)$$

The concentration of RVV is supposed to be constant during each experiment. However, it should be noted that [RVV] differs for the different experiments. The inactivation of II_a and X_a in the presence of ATIII and α_2M is modelled by first order reactions (r_2 , r_8 and r_9). This implies that the concentrations of these inhibitors do not occur in the equations.

The available measurements concern the amidolytic activity, which is expressed as the equivalent amount of thrombin (nM). This means that, in addition to the equations describing the chemistry, an equation for the amidolytic activity should be added. This equation is given in (6.30). It takes into account that the amidolytic activity does not only depend on the activity of thrombin (II_a), but also on the activity of the thrombin inactivated by α_2M ($IIa\alpha_2M$). It is known from [Hem93] that the inactivated form shows an activity of 55.6% of the active thrombin.

In addition to the system of nine differential equations (6.21)-(6.29), we need the same number of initial conditions. At the start ($t = 0$), the initial concentrations of all state variables are zero, except for [PL], [II], [V] and [X].

6.2.5 Parameter estimation and model validation

The system of equations (6.21)-(6.39) contains 13 reaction constants. None of these constants nor the initial concentrations of the coagulation factors [II], [V] and [X] are known exactly, but they are assumed to be constant for each experiment. These 13 reaction constants, plus the three unknown initial conditions, are the quantities we want

to determine; the unknown parameters. We summarise these parameters in Table 6.6. From the current literature we know upper and lower bounds for the concentrations of the clotting factors in normal plasma: i.e. $[750nM, 2200nM]$ for II, $[10nM, 30nM]$ for V and $[70nM, 200nM]$ for X.

The parameters are estimated in such a way that the model responses fit the measurements in a least squares sense. Besides the estimates, confidence regions for the parameters are derived. For more details about the numerical solution of the model equations, minimisation of the least squares criterion, and the confidence regions, the reader is referred to Chapter 1.

To get more insight in our process of model discrimination, we compare each of the four options, (6.17)-(6.20), in combination with the reactions r_2 to r_9 from Figure 6.6 as they are described in (6.32)-(6.39). The expressions for r_2 to r_9 are obtained by a similar process of selection and validation as we will describe below.

Under the assumption of (6.32)-(6.39) we immediately reject option (6.20), because it implies that RVV has no influence on the reaction scheme, which is not in agreement with the experiments.

Under the assumption of (6.32)-(6.39), with one of the options (6.17), (6.18) or (6.19) we compare the corresponding model performances shown in Table 6.5. From this table

r_1	m	df	$S(\hat{\theta})$
(6.17)	16	104	6287×10^3
(6.18)	15	105	7020×10^4
(6.19)	15	105	1005×10^5

Table 6.5: Comparison for the three remaining options (6.17), (6.18) and (6.19). We show the number of parameters (m), the degrees of freedom (df= $N - m$: the number of measurements minus the number of parameters) and the least squares sum ($S(\hat{\theta})$).

it is obvious that the first alternative performs better than the other two, if we take only $S(\hat{\theta})$ into account. In order to decide if one model performs *significantly* better than another, we use the F-ratio test (see Appendix 1.C). To apply this test to the three remaining options for r_1 , we take the reaction scheme from Figure 6.6 and r_2 to r_9 as in (6.32)-(6.39). The relevant data for the F-ratio test are given in Table 6.5. The test of a significant difference between (6.17) and (6.18) consists of constructing a super-model with:

$$r_1 = \left(\frac{kcat_X}{km_X + [X]} + kk_1 \right) [X] \cdot [RVV] . \quad (6.40)$$

The residual sum in case of the super-model is equal to 6091, which is needed to compute

the quantities (cf. (1.36)):

$$X = \frac{(6287 - 6091)/1}{6091/103} = 3.314 , \quad (6.41)$$

in order to compare (6.40) with (6.17), and

$$Y = \frac{(70201 - 6091)/2}{6091/103} = 542.056 , \quad (6.42)$$

to do the same for (6.40) and (6.18). We need to compare X with the upper quantile $\mathcal{F}_{0.05}(1, 103) = 3.93$ and Y with the upper quantile $\mathcal{F}_{0.05}(2, 103) = 3.08$. The bound for Y is exceeded which means that the model with (6.17) accounts significantly better for the phenomena observed. Therefore, r_1 from (6.18) is rejected. Similarly (6.19) is rejected, because it performs even poorer, as can be seen from Table 6.5.

Also, the other models which are derived from alternative schemes described in the Figures 6.7, 6.8 and 6.9, have been tested. All these alternatives give rise to models with more state variables and more parameters. However, following the same strategy none of them turned out to perform significantly better.

6.2.6 Results

An initial estimate for the parameters consists of an educated guess from the existing biochemical literature ([Hem93] and references therein). These initial values are given in Table 6.6. The final estimates, and the corresponding confidence regions are also listed in this table. For details on the statistics, the reader is referred to Section 1.6. The sum of squared residuals for the initial estimates was 2.40×10^7 , after minimisation it was reduced to 6.287×10^3 .

The measurements (120 in total and 30 for each experiment) and the model responses for the final estimates of the parameters are given in Figure 6.10. The plots show a very acceptable fit between the computed and measured values, i.e. a fit within the measurement accuracy, which means that the model gives a sufficiently accurate description of the measured quantities.

The independent and dependent confidence regions as they are listed in the fourth and fifth column of Table 6.6 show that by far not all the parameters can be estimated within reasonable accuracy. From the singular value decomposition of the covariance matrix of the parameters (see Sections 1.5 and 1.6), we can deduce that with the current model and the available measurements 5 parameters (or combinations of parameters) can be estimated with acceptable accuracy. By making use of other chromogenic substrates, additional measurements for V_a and X_a can be obtained in order to estimate more parameters more accurately.

The parameter km_2 tends to become small during the parameter estimation procedure and the idea came up to replace the corresponding reaction, r_7 (cf. (6.37)), with $kk_5 \cdot [Xa]$, in order to reduce the number of parameters by one. The corresponding model gave

parameter	initial est. (θ_{ini})	final est. ($\hat{\theta}$)	independent confidence regions ($\Delta^I \theta$)	dependent confidence regions ($\Delta^D \theta$)
$kcat_X$	5.00×10^3	2.391×10^2	5.301×10^3	1.963×10^1
km_X	4.00×10^2	2.365×10^1	5.776×10^2	6.335×10^0
ki_{Xa}	2.50×10^{-1}	4.531×10^0	1.408×10^1	3.667×10^{-1}
k_{PT}	1.00×10^{-1}	1.229×10^2	3.117×10^5	4.152×10^1
k_{PL}	1.00×10^1	8.014×10^2	2.032×10^6	2.711×10^2
$kcat_V$	1.40×10^1	7.844×10^0	2.166×10^3	1.862×10^0
km_V	7.20×10^1	1.497×10^2	4.261×10^4	3.666×10^1
$kcat_{II}$	2.00×10^3	4.387×10^1	8.678×10^2	2.956×10^0
km_{II}	2.10×10^2	6.225×10^1	2.147×10^2	2.073×10^1
$kcat_2$	2.30×10^0	1.240×10^1	2.596×10^2	9.150×10^{-1}
km_2	5.80×10^1	6.148×10^{-2}	2.937×10^1	1.630×10^1
$ki_{IIaATIII}$	1.30×10^0	7.859×10^{-1}	5.794×10^{-1}	4.423×10^{-2}
$ki_{IIa\alpha_2M}$	1.50×10^0	1.762×10^{-1}	4.611×10^{-2}	2.673×10^{-2}
X_{ini}	1.33×10^2	8.125×10^1	1.729×10^3	7.556×10^0
V_{ini}	1.67×10^1	6.712×10^0	1.663×10^2	5.821×10^{-1}
II_{ini}	1.33×10^3	5.093×10^2	2.677×10^2	2.112×10^1
$S(\theta)$	2.40×10^7	6.287×10^3		

Table 6.6: Initial guess and final estimates for the parameters and their confidence regions.

negative results for the concentration of factor II, which is a consequence of adapting r_7 (the inequality $[II] \gg km_2$ did not hold on the whole time interval), and was therefore rejected.

The term r_7 is inevitable, because without this term the production of thrombin will not even start. This can be seen from the reaction scheme of Figure 6.6 and the fact that the initial concentrations of II_a and V_a are zero. Before the start of the experiments the expectation of the biochemists was that the activation of prothrombin (II) would be mainly performed by prothrombinase (PT) and that the contribution of X_a would be marginal here. In other words: r_7 would be small compared to r_6 and therefore (after initiating the reaction) could be neglected after a few seconds. By investigating the separate contributions to the thrombin production for r_6 and r_7 during the simulations, we found that the contribution of r_7 is about 50% of the production by r_6 and therefore not negligible. This conclusion should, however, be strictly limited to the case of RVV as a factor X activator and be extrapolated to other experimental setups.

Although the results of Table 6.6 may look poor with respect to the confidence regions, it appears that with the current data we were able to discriminate between many models in a systematic way and to come up with a model which fits the observations satisfactorily.

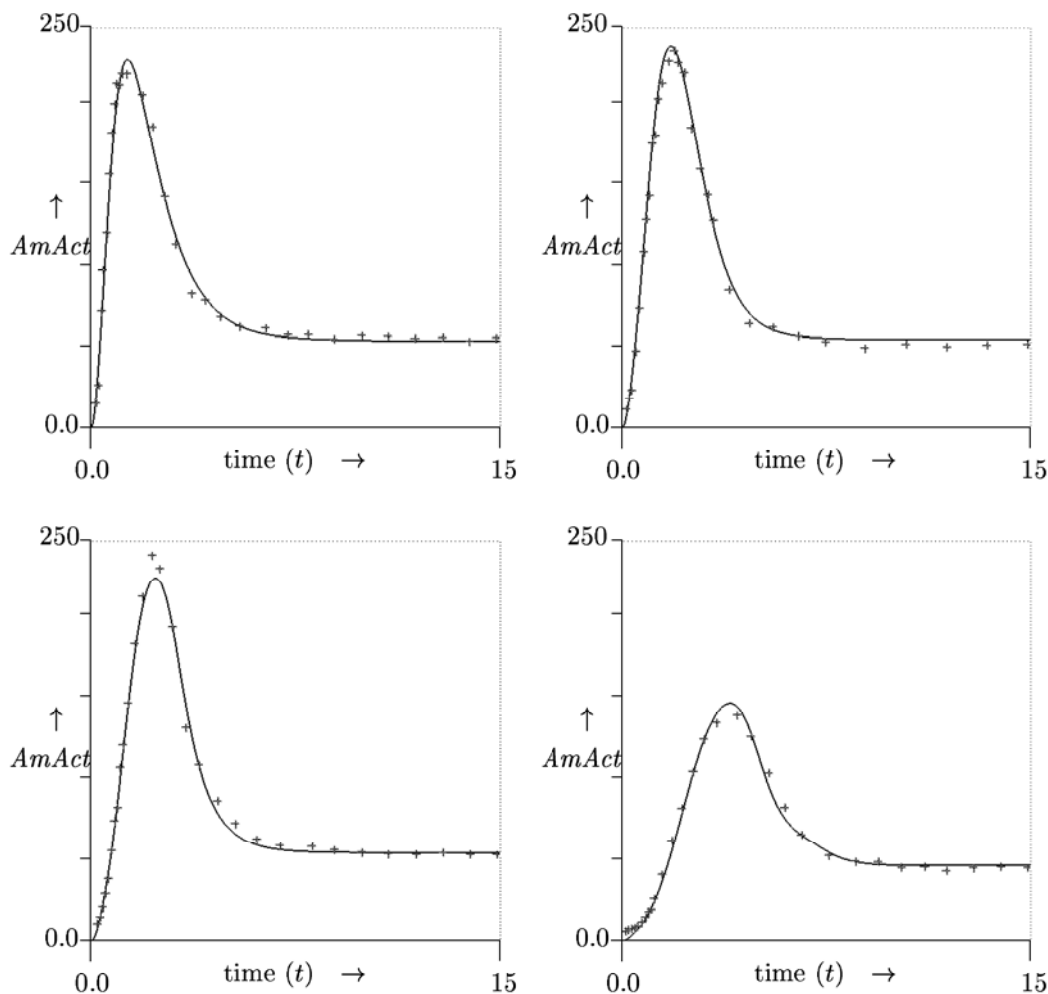


Figure 6.10: Plots of the measurements ('+') and the model responses for the final estimates of the parameters from Table 6.6 over the 4 experiments with decreasing concentrations of RVV.

6.2.7 Conclusions

In this study we compare a number of possible reaction schemes which describe part of the blood coagulation mechanism. For each scheme mathematical model equations have been derived and parameters have been estimated in order to obtain a best fit with a set of experimental data. Depending on the complexity of the model, and the quality

of the fit, judged by the statistical criteria, we were able to discriminate between many candidate models. The final model is compact, meets the established knowledge in the field and fits the measurements satisfactorily. A large number of more compact models were rejected on the account of the measurements. More sophisticated models were rejected because the increase of complexity did not account for a sufficient improvement of the fit.

With the final model selected not only its parameter estimates are presented, which are optimal in a least squares sense with respect to the available data, but also the corresponding confidence regions. Additional experiments can make the confidence regions smaller, while on the other hand they may also lead to a more complex model in favour of one of the alternatives which had to be rejected in this study.

In this sense the presented model can be a good starting point for ongoing research and may show its value when more experimental data are available.

Acknowledgement:

P. Devilee (Department of Biochemistry, University Maastricht) is gratefully acknowledged for the numerous discussions, his work on preparing the data and his patience to deal with a layman experimenting in the lab.

6.3 Production by plant cells in suspension[†]

Symphytum officinale L. cells were grown in Erlenmeyer flasks at four different temperatures: 15, 20, 25 and 30°C. A mathematical model of the culture growth is presented. The intracellular and extracellular products are considered in separate equations. An interrelation between fresh weight, dry weight and viability is considered in the balances. The model includes a description of the changes in time of wet and dry biomass, cell viability, substrate concentration and polysaccharide concentration, both intra- and extracellular. The model was tested by fitting the numerical results to the data obtained.

6.3.1 Introduction

Cell suspension cultures are of industrial interest because of their potential for the controlled synthesis of high price natural products that are found only in plants, and are usually obtained by extraction from the whole plant tissue. There are only few commercial processes for the production of plant cell metabolites in suspension culture. One of the obstacles in the scale-up of such processes is the lack of adequate kinetic descriptions of the phenomena involved in mathematical terms. Mathematical models are useful for predicting the behaviour and determining the optimum operating conditions for a process with a minimum of experiments on large scale, which are very expensive. For the case of a batch process, a mathematical model should be able to predict the time-course of the culture in the bioreactor. Such models have proved to be very successful in microbiological processes. The models proposed range from very simple unstructured ones, which are able to predict only the variations of biomass in time [Fra89, MA95] to complex structured models describing the variation of many of the components in the cell, their interaction and the formation of products [SD83].

The description of plant cells in suspended cultures presents some particularities which complicate the description of the system in mathematical terms. One of them is the existence of nonviable cells in proportions much higher than in usual microbial cultures. A satisfactory description must therefore include the balance of viable and nonviable cells in the bioreactor, as well as the product formation. Several structural models have been proposed for the description of plant cells [Pol86, Wei89]. Bailey and Nicholson [BN89] proposed the ratio of fresh weight to viable dry weight to express the susceptibility of cells to shear stress and to relate the loss of cell viability to this ratio. They fitted their model to the production of alkaloids by cells of *Catharantus roseus*.

Some polysaccharides have therapeutic properties [GR86, Neu90] and are an important commodity in the food industry [WB73]. There are some reports of polysaccharide (PS) accumulation in liquid media of plant cells in suspension [BKMA74, HPD87]. The extracellular polysaccharide (EPS) is either similar to [BKMA74] or different from the

[†] This section results from joint work with Ruha Glicklis and Jose Merchuk (Program of Biotechnology, Department of Chemical Engineering, Ben Gurion University of the Negev, Beer Sheva, Israel) and has been submitted in an almost identical form to *Biotechnology & Bioengineering*.

cell wall PS [YS77]. Differences in the composition of EPS were found among cells of different species [BKMA74]. Cells of *Phleum pratense* were shown to secrete fructans to the medium [HPD87]. Becker et al. [BHA64] reported that EPS production paralleled the growth of cells of *Acer pseudoplatanus* in batch cycle. As far as we know, no one has characterised further the kinetics of PS production in cell suspensions.

In this case study, a mathematical model for PS production in a cell suspension of *Symphytum officinale* L. is presented, making use of the elements of expansion and lysis phase as proposed by Bailey and Nicholson [BN89]. The intracellular and extracellular products are considered in separate equations. Furthermore, the interrelation between fresh weight, dry weight and viability is considered in the balances. The unknown parameters of the mathematical model were evaluated by fitting its results to experimental data obtained in cultures grown in Erlenmeyer flasks (at four different temperatures). The state variables of the mathematical model include the measured quantities (i.e. concentrations of substrate, fresh and dry weight, intracellular and extracellular PS, and cell viability).

6.3.2 Materials and methods

The *S. officinale* cell suspension was initiated from callus and was grown in MS medium [MS62], supplemented with 0.2 mg/L 2,4-Dichlorophenoxyacetic acid, 0.2 mg/L kinetin, 100 mg/L *p*-chlorophenoxyacetic acid, and 30 g/L sucrose. The pH was adjusted to 5.8. Cultures were subcultured every 17 days using a 10% (v/v) inoculum and maintained in 250 mL Erlenmeyer flasks containing 100 mL. Cultures were incubated in the dark at 25°C on a shaker at 150 rpm.

Observation under a microscope of the samples taken showed that during the first stage of the culture most of the population were single separate cells, with some pairs and trios. After the tenth day the number of those formations increased and some clumps of a slightly larger size could be seen as well, of the order of ten cells. Some chains of four-five cells could be seen. Nevertheless, most of the cells stayed single.

Every 2-3 days, cells were harvested from three Erlenmeyer flasks and filtered by buchner funnel. The filtrate was kept for sugar and PS determination. After determining the fresh weight, viability was determined by flourocein diacetate dying [Wid72]. Dry weight was determined by placing samples in an oven and maintained at 70°C for 10 hours.

For determining intracellular polysaccharide (P_1), dry cells were ground with a pestle and mortar and extracted first by boiling in de-ionised water for 10 min and then by stirring for 3 h at room temperature. Cell debris was removed by centrifugation at 1000 rpm. The supernatant as well as the filtrate (extracellular fluid) of each fresh cell harvest, for the extracellular polysaccharide (P_2) determination, were frozen at -18°C and then dried by lyphilysation. Tanins were removed by 2% $PbSO_4$ and after centrifugation at 5000 rpm extra lead was removed by 1% oxalic acid, followed by another centrifugation at 5000 rpm. The supernatant was frozen and lyphilysed once more. The dry material

was dissolved in 2 *mL* of de-ionised water and polysaccharides were precipitated in 10% (v/v) ethanol after storage for overnight at 4°C. Pellets were lyophilised and weighted for the determination of intracellular and extracellular polysaccharides.

Sucrose concentration was evaluated by colorimetric measurement of reducing sugars after hydrolysis [CK86].

6.3.3 Model development

Conceptual Model

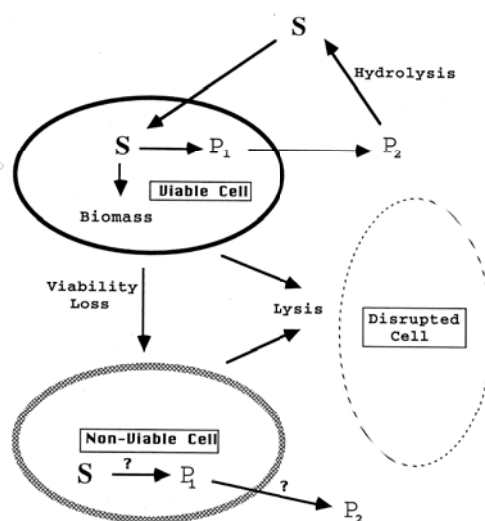


Figure 6.11: Schematic representation of the model assumed for cell growth and polysaccharide production.

The structured kinetic model initially proposed, accounting for growth, cell expansion and lysis, polysaccharide formation, secretion and hydrolysis in the medium, is shown schematically in Figure 6.11. Viable cells consume the substrate present in the medium, and may either produce new viable cells, transform into nonviable cells or undergo lysis. The nonviable cells are generated from the viable cells, and disappear due to lysis. It is assumed that only viable cells produce PS. Furthermore, substrate consumption for maintenance is neglected, and therefore only viable cells consume sucrose, for both

biomass generation and product synthesis. Viability (V) is defined as the fraction of cell dry weight that is viable (FDA staining), and takes values between 0 and 1. The model considers that the cell lysis causes decay in nonviable dry weight and viable dry weight at different rates, both being expressed by first order kinetics with constants k_d , k'_d . Polysaccharide is synthesised inside the viable cells and is secreted to the medium, where it is partially hydrolysed. The production rate of polysaccharide may be proportional to the growth rate (growth associated product) or independent of it. Both possibilities were considered and it was concluded that in the present case the polysaccharide production rate is growth associated.

Biomass balance:

The balance on viable dry weight is written as follows:

$$\frac{dX_{vd}}{dt} = \left[\frac{\mu_{max}S}{k_S + S} - k_i(X_f/X_{vd})^2 - k_d \right] X_{vd} . \quad (6.43)$$

It considers that dry mass is produced at a specific growth rate which can be expressed by a Monod type kinetics with constants μ_{max} and k_S . The second term in Eq. (6.43) represents the transfer of viable cells to nonviable cells at a rate which is first order in viable cell concentration, and second order in the following ratio defined by Bailey and Nicholson [BN89]:

$$\chi = X_f/X_{vd} , \quad (6.44)$$

which is supposed to be an indication of cell size expansion, assuming all cells are of the same dry weight. The mentioned authors found that this kinetic form gave the best fit for their data. The same was found for the data presented here. The third term in Eq. (6.43) represents the consumption of the viable dry weight by lysis at a rate which is first order in the viable dry weight.

Nonviable dry weight is generated from the viable dry weight, as shown in Eq. (6.43), and is lost by lysis with first order kinetics, which yields:

$$\frac{dX_{nd}}{dt} = k_i(X_f/X_{vd})^2 X_{vd} - k'_d X_{nd} . \quad (6.45)$$

The total dry weight is the sum of the viable dry weight (X_{vd}) and nonviable dry weight (X_{nd}):

$$\frac{dX_d}{dt} = \left[\frac{\mu_{max}SV}{k_S + S} - k_dV - k'_d(1 - V) \right] X_d , \quad (6.46)$$

where viability is defined as the ratio of viable dry weight and dry weight:

$$V = X_{vd}/X_d . \quad (6.47)$$

Substrate balance:

Given the initial composition of the medium, sucrose is the limiting substrate. It has been suggested [UIFN74], that immobilised invertase on the cell wall catalyses the hydrolysis of sucrose into glucose and fructose, which are absorbed into the cell.

The product synthesis rate can be considered growth associated. As a consequence, Eq. (6.48) describes the conversion of S into dry weight with a constant yield Y_{sx} and into polysaccharide according to a Monod type kinetics. It is assumed that no sucrose is consumed for maintenance:

$$\frac{dS}{dt} = -\frac{\mu_{max} S X_{vd}}{(k_S + S)} \frac{1}{Y_{sx}} + \frac{k_4 P_2 X_{vd}}{k_p + P_2 + S^2/k_c} . \quad (6.48)$$

The first term in the rate equation is the consumption for growth appearing in Eq. (6.43), divided by the yield, and the second term represents the production of S by hydrolysis of polysaccharide product in the medium, which will be justified in the next paragraphs. There is no need to account here for the consumption of S for P_1 synthesis, since in this growth-associated scheme Y_{sx} accounts for all substrate consumption.

Intracellular polysaccharide balance:

It is assumed that the P_1 concentration results from a balance between formation rate and the secretion to the medium. Polysaccharide concentration inside the cell will increase with a rate that is proportional to the growth rate of biomass. Assuming that substrate transfer into the cell is not limiting, so that S concentration inside the cell is the same as in the bulk of the medium:

$$r_{(PS \text{ synthesis})} = \frac{Y_{sp}}{Y_{sx}} \frac{\mu_{max} S X_{vd}}{k_S + S} . \quad (6.49)$$

Equation (6.49) is given in mass of polysaccharide produced referred to the whole volume of the culture.

The rate of secretion of polysaccharide to the medium was assumed to be proportional to two factors: 1) To the difference between the actual concentration of polysaccharide inside the cell and its concentration in the medium. 2) To the interfacial area of the cells. Assuming that the interfacial area is proportional to the fresh cell concentration, X_f (which will be close to reality if the distribution of cell aggregates is constant), and that it is proportional to the reciprocal of its size (which is represented by χ , Eq. (6.44)), the rate of polysaccharide secretion can be expressed as follows:

$$\begin{aligned} \frac{dP_1}{dt} = & \left[\frac{Y_{sp}}{Y_{sx}} \frac{\mu_{max} S}{k_S + S} - ka(1 - P_2/P_1) \right] X_{vd} + \\ & \left[\frac{Y'_{sp}}{Y_{sx}} \frac{\mu_{max} S}{k'_S + S} - ka'(1 - P_2/P_1) \right] X_{nd} , \end{aligned} \quad (6.50)$$

where the second term of the right-hand member represents the parallel phenomenon in the nonviable cells.

Extracellular polysaccharide balance:

The secretion process is responsible for the transfer of polysaccharide from the cell to the medium. The first term in Eq. (6.51) represents the polysaccharide transferred from the cells. It is assumed that the accumulated polysaccharide is partially hydrolysed in the medium in order to supply glucose for maintenance. This parallels the phenomenon that occurs in the intact plant where polysaccharide is hydrolysed by polysaccharide hydrolase enzyme to rebuild the plant in the growth season [EJ68]. When the sucrose level is too high, the fraction hydrolysis is inhibited. The second term in Eq. (6.51) represents the polysaccharide hydrolysis with constants k_4 , k_p and k_c for growth, saturation and inhibition respectively. This type of inhibition kinetic had been suggested by Andrews [And68] for microbial cultures. Consequently we have

$$\frac{dP_2}{dt} = ka(1 - P_2/P_1)X_{vd} + ka'(1 - P_2/P_1)X_{nd} - \frac{k_4 P_2 X_{vd}}{k_p + P_2 + S^2/k_c} . \quad (6.51)$$

Fresh biomass balance:

Increase in fresh weight is due to both cell growth and expansion. Knowledge of X_f is not required by the equations modeling the growth of the cells, but is needed to model product synthesis and excretion. The experiments run in our laboratory showed that for all the temperatures and both in flasks and bioreactor, X_d would level off after certain period, and then decrease. This coincides with a sharper decrease in viability. The fresh weight, on the other hand, keeps increasing continuously. This seems to indicate that a relationship exists between cell viability and expansion.

It was found that the expression used by Bailey and Nicholson [BN89]:

$$\frac{dX_f}{dt} = Z V X_d + \frac{dV}{dt} X_d \chi , \quad (6.52)$$

where Z is a constant, allowed a satisfactory fit of our experimental data, and was used to provide the link between dry and fresh biomass. After some algebraic manipulation, Eq. (6.60), as shown later, is obtained.

6.3.4 Experimental results and final model confirmation

The model as derived for this case study and the corresponding measurements built a parameter estimation problem which can be solved by the techniques as introduced in Chapter 1 to 5. The process of parameter estimation for optimal fitting of the experimental results did not only render the numerical values which allow the mathematical modeling the culture growth. It was also instrumental in evaluating the proposed model. The first conclusion that could be obtained from the mathematical model was that the production of polysaccharides was a growth associated process. When the optimal value obtained for a parameter was very small, an F-ratio test (see Appendix 1.C) was performed in order to decide whether the parameter could be omitted. If a parameter can be omitted without a significantly poorer fit, it was considered as an indication that the

model itself had to be modified in this respect, and that the parameter had to be eliminated from the formulation. All the Monod type kinetics proposed initially were finally replaced by first order kinetics, without affecting the fit. In the case of the growth rate an actual first order specific reaction rate with respect to both X_{vd} and S can be defined, with a maximal growth rate μ_{maxa} . Since $k_S \gg S$, it is related to the parameters of the initial model as follows:

$$\mu_{maxa} = \mu_{max}/k_S \approx \mu_{max}/(k_S + S) . \quad (6.53)$$

Similarly, for polysaccharide hydrolysis to S , since $k_p \gg [P_2 + S^2/k_c]$:

$$k_{4a} = k_4/k_p \approx k_4/(k_p + P_2 + S^2/k_c) . \quad (6.54)$$

The value of the decay constant for viable cells, k_d , was found to be negligible at low temperatures. As a consequence the term of consumption of cells could be eliminated in the balance of viable cells in Eq. (6.43). This is not just a simplification of the mathematical formulation, but –more importantly– an indication on the mechanism of the process. In particular, this indicates that viable cells are much more resistant to shear stress and other environmental damages, and mainly nonviable cells undergo lysis in the culture at low temperature. In addition to this, the results of the parameter optimisation done with the model suggest that the nonviable cells do not take part in the production of polysaccharide product ($Y'_{sp} = 0$; $ka' = 0$).

The final model, after all modifications, can be formulated as:

$$\frac{dX_{vd}}{dt} = [(\mu_{maxa}S) - k_i(X_f/X_{vd})^2 - k_d] X_{vd} , \quad (6.55)$$

$$\frac{dX_{nd}}{dt} = k_i X_{vd} (X_f/X_{vd})^2 , \quad (6.56)$$

$$\frac{dS}{dt} = [k_{4a}P_2 - (\mu_{maxa}S)/Y_{sx}] X_{vd} , \quad (6.57)$$

$$\frac{dP_1}{dt} = [(\mu_{maxa}S)(Y_{sp}/Y_{sx}) - ka(1 - P_2/P_1)] X_{vd} , \quad (6.58)$$

$$\frac{dP_2}{dt} = [ka(1 - P_2/P_1) - k_{4a}P_2] X_{vd} , \quad (6.59)$$

$$\begin{aligned} \frac{dX_f}{dt} = & Z X_{vd} + [X_{nd}(\mu_{maxa}S X_{vd} - k_i X_{vd} (X_f/X_{vd})^2 - k_d X_{vd}) - \\ & X_{vd}(k_i X_{vd} (X_f/X_{vd})^2 - k'_d X_{nd})] X_f / (X_d X_{vd}) , \end{aligned} \quad (6.60)$$

$$V = X_{vd}/X_d , \quad (6.61)$$

$$X_d = X_{vd} + X_{nd} . \quad (6.62)$$

Measurements for the state variables: S , X_f , X_d , P_1 , P_2 and V , taken at different temperatures are available. The parameters to be estimated are: μ_{maxa} , k_i , k_d , ka , k_{4a} , Y_{sx} , Z and Y_{sp} . As fitness criterion we took the sum of squared discrepancies (recall (1.4)). The weights are proportional to the accuracy of the measurements, as indicated by

the error bars in Figures 6.12-6.15¹. The choice for the weights will be highlighted at the end of this section. The optimal parameter values at each temperature, the independent and dependent confidence regions (cf. (1.25) and (1.26), respectively) for $\alpha = 0.05$ are shown in the Tables 6.8-6.11.

Figure 6.12 shows model response curves and experimental data for *S. officinale* culture growth at 30°C. The top part of Figure 6.12 shows the profiles of sucrose S , and fresh biomass X_f . For S and X_f the mean of three experimental measurements and the standard deviations are shown. The curves correspond to the model responses, evaluated for the optimal, i.e. estimated values of the parameters as shown in Table 6.8. The descent of the sucrose concentration and the increase of the wet biomass are closely fit by the model.

The middle part of Figure 6.12 displays the experimental and calculated profiles of dry biomass concentration X_d , the concentration of viable dry biomass X_{vd} and the concentration of nonviable dry biomass X_{nd} , at 30°C. A strong decrease in the concentration of viable dry biomass is seen after 20 days. As will be seen in the following graphs, this effect diminishes at lower temperatures. The model follows this trend with a satisfactory fit. The viability, Eq. (6.47), is represented in the bottom graph of Figure 6.12, together with the concentration of both intracellular and extracellular polysaccharide. The figure shows that a decrease in both P_1 and P_2 is observed, starting approximately at the same time as the decrease in viability.

The graphs of Figure 6.13 display similar results of experiments run at 25°C and the corresponding calculated profiles. In a similar way the results corresponding to 20°C and 15°C are shown in the graphs of the Figure 6.14 and 6.15, respectively. In all of these cases the mathematical model is able to represent adequately the experimental results. An exception is the curve representing the substrate concentration S in the runs at 15°C, in the first part of the experiment (till approximately 15 days). This seems to be due to an experimental error, since there is no reason for an actual increase of sucrose in the medium. The optimal values found for the constants of the kinetic model are shown, together with the corresponding statistical data, in Tables 6.9 to 6.11 for 25°C, 20°C and 15°C, respectively.

Comparing the profiles of the state variables, it can be seen that as the culture temperature decreases, the rate of decrease of sucrose decreases. The final concentration of wet biomass is higher at higher temperature.

The decrease in cell viability is strongly related to temperature. The higher the temperature, the larger the loss in viability. At 15°C almost no losses are detected during the culture period. This can be appreciated not only in the graph of viability, but also in the profiles of biomass.

The decision with respect to the weights was taken by the experimentalists on the basis of the error bars and their knowledge on the experimental setup and equipment used. The weights they came up with and are used throughout this case study, are given

¹For X_{nd} and X_{vd} no measurements are available. The position of corresponding markers in the figures is based on (6.61) and (6.62), but does not influence the estimation process.

in Table 6.7.

We used both these weights and estimated weights as described in Section 3.4. We still assume the measurement errors to be independent. The estimated weights are given in Table 6.7, where it can be seen that, except for an unimportant factor of about 10, the weights are close to each other. The parameters, estimated in this way, show a

Component	Exp. weight	MLE weight
S	0.03	0.39
X_f	0.01	0.11
X_d	0.1	2.3
P_1	3.0	28.3
P_2	3.0	62.6
V	3.0	30.1

Table 6.7: Weights for each measured component derived by the experimentalists and calculated as in Section (3.4).

close correspondence to the estimates we had already (within the dependent confidence regions) and the changes in plots of the response variables were marginal. The alternative approach needed 11 iterations to converge, where the approach from Chapter 1 needed 10 iterations.

6.3.5 Conclusions

Comparison of experimental data of growth *Symphytum officinale* L. cells in Erlenmeyer flasks at four different temperatures showed excellent agreement with a mathematical model proposed. The model describes changes in time of wet and dry biomass, cell viability, substrate concentration and PS concentration, both intra- and extracellular. The model assumed that the production of polysaccharides is growth associated. Furthermore, the analysis of the mathematical model led to the conclusion that the nonviable cells are not active in product formation, and that mainly nonviable cells undergo lysis during the growth of the culture.

The model as presented is a very useful tool for simulation of growth of plant cells cultures and polysaccharide synthesis rate. The comparison of a weighted least squares approach with a MLE approach with unknown weights showed a close agreement. This means that, if no a priori knowledge about the measurement error would have been used, it would not have affected the final answers significantly. In other words: the a priori error assumption matches the a posteriori error structure.

Acknowledgement: Mrs. Shivta Vencart is sincerely acknowledged for the invaluable assistance in the experimental work.

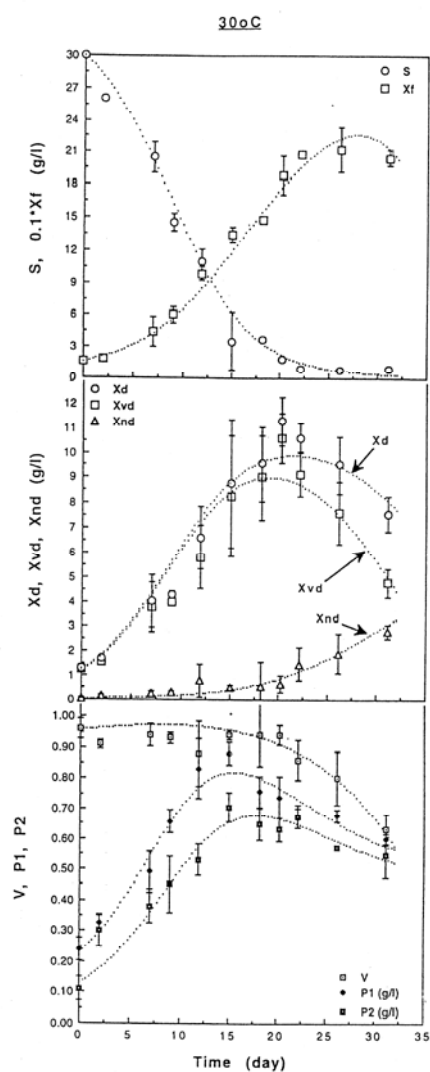


Figure 6.12: Measurements and optimal fit (cf. Table 6.8) of sucrose S and fresh biomass X_f (top), dry biomass X_d , viable dry biomass X_{vd} and nonviable dry biomass X_{nd} (middle), cell viability V , intracellular polysaccharide concentration P_1 and extracellular polysaccharide concentration P_2 (bottom) at 30°C.

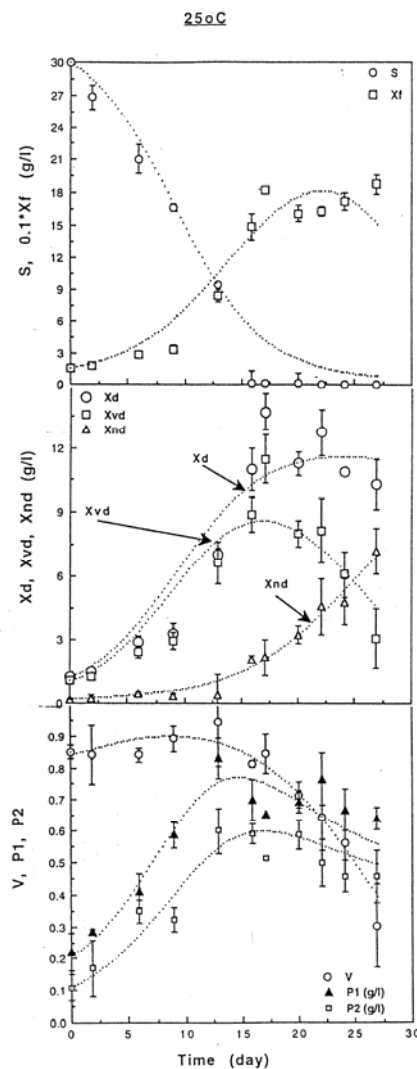


Figure 6.13: Measurements and optimal fit (cf. Table 6.9) of sucrose S and fresh biomass X_f (top), dry biomass X_d , viable dry biomass X_{vd} and nonviable dry biomass X_{nd} (middle), cell viability V , intracellular polysaccharide concentration P_1 and extracellular polysaccharide concentration P_2 (bottom) at 25°C.

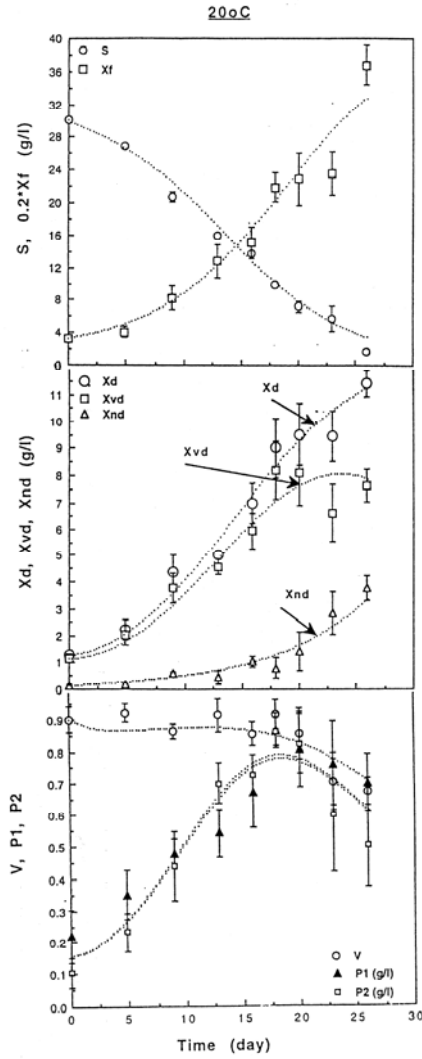


Figure 6.14: Measurements and optimal fit (cf. Table 6.10) of sucrose S and fresh biomass X_f (top), dry biomass X_d , viable dry biomass X_{vd} and nonviable dry biomass X_{nd} (middle), cell viability V , intracellular polysaccharide concentration P_1 and extracellular polysaccharide concentration P_2 (bottom) at 20°C.

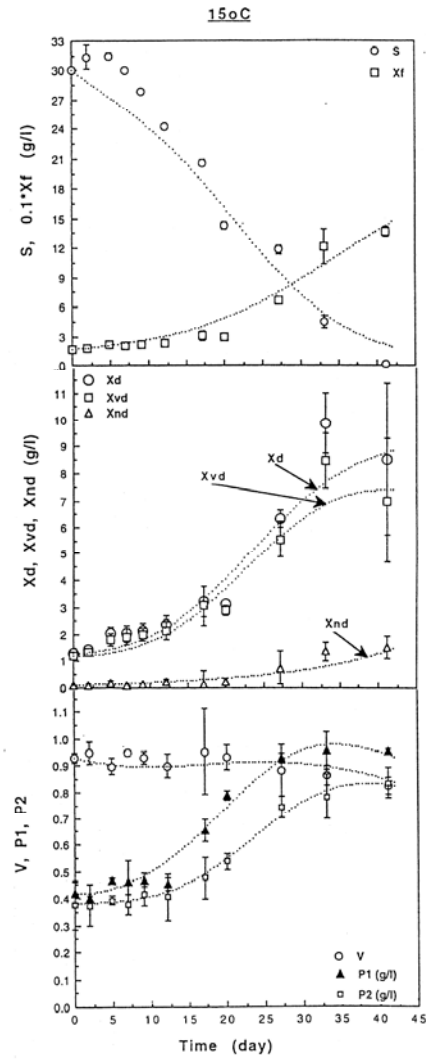


Figure 6.15: Measurements and optimal fit (cf. Table 6.11) of sucrose S and fresh biomass X_f (top), dry biomass X_d , viable dry biomass X_{vd} and nonviable dry biomass X_{nd} (middle), cell viability V , intracellular polysaccharide concentration P_1 and extracellular polysaccharide concentration P_2 (bottom) at 15°C.

Nomenclature used in Section 6.3

ka	Secretion constant in the viable cells ($gP * gX^{-1} * day^{-1}$).
ka'	Secretion constant in the nonviable cells ($gP * gX^{-1} * day^{-1}$).
k_c	Inhibition constant ($gS^2 * gP^{-1} * L^{-1}$).
k_d	Decay constant (day^{-1}).
k_i	Mortality constant (day^{-1}).
k_p	Product hydrolysis saturation constant ($gP * L^{-1}$).
k_S	Growth saturation constant ($gS * L^{-1}$).
k_4	Specific product hydrolysis rate ($gS * gX^{-1} * day^{-1}$).
k_{4a}	Specific product hydrolysis rate, final model ($L * gX^{-1} * day^{-1}$).
P_1	Intracellular polysaccharide concentration per volume of culture ($g * L^{-1}$).
P_2	Extracellular polysaccharide concentration per volume of culture ($g * L^{-1}$).
S	Sucrose concentration ($g * L^{-1}$).
V	Viability (-).
X_d	Dry weight ($g * L^{-1}$).
X_{nd}	Nonviable dry weight ($g * L^{-1}$).
X_{vd}	Viable dry weight ($g * L^{-1}$).
X_f	Fresh weight ($g * L^{-1}$).
Y_{sx}	Biomass yield ($gX_d * gS^{-1}$).
Y_{sp}	Production yield in the viable cells ($gP * gS^{-1}$).
Y'_{sp}	Production yield in the nonviable cells ($gP * gS^{-1}$).
Z	Expansion coefficient (day^{-1}).
μ_{max}	Specific growth rate (day^{-1}).
μ_{maxa}	Specific growth rate, final model ($L * gS^{-1} * day^{-1}$).
χ	Size parameter, given by Equation (6.44) (-).

Tables

Parameters	Units	Value	Ind.conf.reg.	Dep.conf.reg.
$\mu_{maxa} \times 10^3$	$L \cdot gS^{-1} \cdot day^{-1}$	8.227	2.251	0.724
$k_i \times 10^5$	day^{-1}	3.6	2.1	1.0
$k_d \times 10^3$	day^{-1}	23.96	42.24	6.277
$ka \times 10^2$	$gP \cdot gX^{-1} \cdot day^{-1}$	4.217	2.943	1.824
$k_{4a} \times 10^3$	$L \cdot gX^{-1} \cdot day^{-1}$	10.33	7.361	1.706
Y_{sx}	$gX_d \cdot gS^{-1}$	0.398	0.232	0.036
Z	day^{-1}	1.473	0.363	0.196
Y_{sp}	$gP \cdot gS^{-1}$	0.062	0.023	0.004

Table 6.8: Optimal parameters for Erlenmeyer flasks culture at 30°C.

Parameters	Units	Value	Ind.conf.reg.	Dep.conf.reg.
$\mu_{maxa} \times 10^3$	$L \cdot gS^{-1} \cdot day^{-1}$	8.093	3.523	1.090
$k_i \times 10^5$	day^{-1}	11.7	7.1	2.8
$k_d \times 10^3$	day^{-1}	3.378	74.91	9.179
$ka \times 10^2$	$gP \cdot gX^{-1} \cdot day^{-1}$	3.930	4.433	2.299
$k_{4a} \times 10^3$	$L \cdot gX^{-1} \cdot day^{-1}$	13.59	15.96	3.40
Y_{sx}	$gX_d \cdot gS^{-1}$	0.372	0.355	0.047
Z	day^{-1}	1.722	0.633	0.304
Y_{sp}	$gP \cdot gS^{-1}$	0.060	0.39	0.062

Table 6.9: Optimal parameters for Erlenmeyer flasks culture at 25°C.

Parameters	Units	Value	Ind.conf.reg.	Dep.conf.reg.
$\mu_{maxa} \times 10^3$	$L \cdot gS^{-1} \cdot day^{-1}$	7.470	6.594	1.884
$k_i \times 10^5$	day^{-1}	12.2	15.6	9.4
$k_d \times 10^3$	day^{-1}	0.000	120.7	21.24
$ka \times 10^2$	$gP \cdot gX^{-1} \cdot day^{-1}$	2.000	0.000	0.000
$k_{4a} \times 10^3$	$L \cdot gX^{-1} \cdot day^{-1}$	1.04	8.87	4.02
Y_{sx}	$gX_d \cdot gS^{-1}$	0.318	0.531	0.106
Z	day^{-1}	1.050	0.699	0.476
Y_{sp}	$gP \cdot gS^{-1}$	0.042	0.022	0.007

Table 6.10: Optimal parameters for Erlenmeyer flasks culture at 20°C.

Parameters	Units	Value	Ind.conf.reg.	Dep.conf.reg.
$\mu_{maxa} \times 10^3$	$L \cdot gS^{-1} \cdot day^{-1}$	2.933	3.757	0.344
$k_i \times 10^5$	day^{-1}	4.5	4.9	2.9
$k_d \times 10^3$	day^{-1}	0.000	78.10	0.841
$ka \times 10^2$	$gP \cdot gX^{-1} \cdot day^{-1}$	1.298	2.086	0.841
$k_{4a} \times 10^3$	$L \cdot gX^{-1} \cdot day^{-1}$	0.000	6.429	1.590
Y_{sx}	$gX_d \cdot gS^{-1}$	0.327	0.582	0.037
Z	day^{-1}	0.675	0.341	0.234
Y_{sp}	$gP \cdot gS^{-1}$	0.042	0.034	0.006

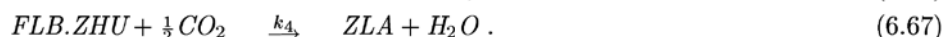
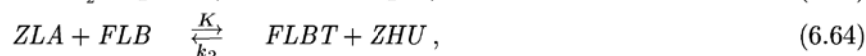
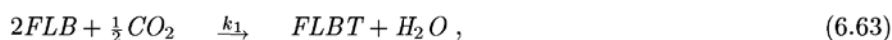
Table 6.11: Optimal parameters for Erlenmeyer flasks culture at 15°C.

6.4 ZLA-kinetics

In this section we discuss a problem which originates from the Akzo Nobel research laboratory in Arnhem, The Netherlands. A slightly different model describing the same chemistry is also a part of the test set for IVP solvers [LSV96]. The names of the chemical compounds are fictitious. Due to the origin of this problem no background on the chemistry is given.

6.4.1 Description of the chemical reactions

In the process under consideration two chemical components, denoted by *FLB* and *ZHU*, are mixed under an inflow of carbon dioxide. The product of interest remaining at the end of the reaction is *ZLA*. The reaction mechanism, as given by Akzo Nobel, reads:



The mechanism of (6.66) is assumed to describe a fast equilibrium:

$$Ks = \frac{[FLB.ZHU]}{[FLB] * [ZHU]}. \quad (6.68)$$

The square brackets denote the concentration of a species in *mol/l*. We identify the concentrations $[FLB]$, $[CO_2]$, $[FLBT]$, $[ZHU]$, $[ZLA]$ and $[FLB.ZHU]$ with the time dependent state variables y_1, \dots, y_6 . The reaction rates to be estimated, k_1 , k_2 , k_3 , k_4 and K , are denoted by the vector θ . The fast equilibrium is taken care of in Section 6.4.2. For this case study we will not focus on the process of model discrimination and validation, therefore we only give the resulting reaction kinetics:

$$r_1 = \theta_1 * y_1^4 * \sqrt{y_2}, \quad (6.69)$$

$$r_2 = \theta_2 * y_3 * y_4, \quad (6.70)$$

$$r_3 = \theta_5 * y_1 * y_5, \quad (6.71)$$

$$r_4 = \theta_3 * y_1 * y_4^2, \quad (6.72)$$

$$r_5 = \theta_4 * y_6^2 * \sqrt{y_2}. \quad (6.73)$$

Besides the above reaction mechanism, there is an inflow of carbon dioxide (given in *ml/min*):

$$F_{in} = 22400 * Vr * kLA * \left(\frac{p(CO_2)}{H} - [CO_2] \right). \quad (6.74)$$

Here we introduced the following abbreviations: Vr : reaction volume; kLA : mass transfer coefficient; H : the Henry constant ($=737 \text{ bar} * l/mol$) and $p(CO_2)$ denotes the partial carbon dioxide pressure. Vr , H and $p(CO_2)$ are a priori known constants; kLA is estimated in the parameter estimation procedure.

6.4.2 Problem description of ZLA-kinetics

Combining the reaction scheme (6.63)-(6.67) with kinetics from (6.69)-(6.73), the evolution of the process is described by the system of differential equations:

$$y_1' = -2r_1 + r_2 - r_3 - r_4, \quad (6.75)$$

$$y_2' = -\frac{1}{2}r_1 - r_4 - \frac{1}{2}r_5 + F_{in}^*, \quad (6.76)$$

$$y_3' = r_1 - r_2 + r_3, \quad (6.77)$$

$$y_4' = -r_2 + r_3 - 2r_4, \quad (6.78)$$

$$y_5' = r_2 - r_3 + r_5, \quad (6.79)$$

$$y_6' = -r_5. \quad (6.80)$$

As a consequence of the fast equilibrium from (6.68), which becomes $Ks = y_6/(y_1 y_4)$, one of the differential equations for either y_1 , y_4 or y_6 can be replaced by an algebraic equation representing this fast equilibrium.

The measurements performed yield data on the inflow of carbon dioxide, cf. (6.74), at a sequence of times. This inflow is given in ml/min and the resulting change of the carbon dioxide concentration due to this inflow, F_{in}^* from (6.76), is given in: $mol/(l * min)$. The relation between these two quantities is:

$$F_{in}^* = \frac{F_{in}}{22400 * Vr}. \quad (6.81)$$

The inflow of carbon dioxide, F_{in} , is described by an additional state variable: y_7 , therefore (6.74) is rewritten as: Combining the equations (6.74) and (6.81) we obtain the algebraic equation:

$$y_7 = 22400 * Vr * kLA * \left(\frac{p(y_2)}{H} - y_2 \right). \quad (6.82)$$

Furthermore, by (6.81) and (6.82) equation (6.76) can be rewritten as:

$$y_2' = -\frac{1}{2}r_1 - r_4 - \frac{1}{2}r_5 + \frac{y_7}{22400 * Vr}. \quad (6.83)$$

The initial concentrations for y_1 , y_4 and y_6 are given to be 0.804, 0.367 and 0.000, respectively, but in view of the fast equilibrium of (6.66) we better take $0.804 - \theta_7$, $0.367 - \theta_7$ and θ_7 as their initial values. Due to the choice of $y_6(0) = \theta_7$ as a free parameter, the equilibrium constant Ks depends on this θ_7 (cf. (6.68)). Taking Ks as a

parameter and deriving the initial concentrations for y_1 , y_4 and y_6 leads to unnecessary complications.

The unknown parameters in the process are the entries of the vector $\theta \in \mathbb{R}^7$:

$$\theta^T = (k_1, k_2, k_3, k_4, K, klA, y_6(0))^T,$$

where all parameters are positive and an additional restriction is given for the last parameter; $\theta_7 \leq 0.367$. An initial guess for the parameters to be estimated was available from Akzo Nobel. The initial values for the state variables are:

$$y(0) = (0.804 - \theta_7, 0.00123, 0, 0.367 - \theta_7, 0, \theta_7, 0)^T.$$

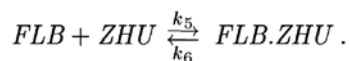
At this point we completed the formulation of a parameter estimation problem.

6.4.3 Parameter estimation results for ZLA-kinetics

The initial model, used as the starting point for the numerical investigation, is given by (6.75), (6.77)-(6.80), (6.82) and (6.83). With this model the parameters k_1 and k_2 tend to zero ($< 10^{-10}$). From an F-ratio test (see Appendix 1.C) it was clear, that, with the available data and this model, k_1 and k_2 can be omitted.

The final model, i.e. the above model after omission of k_1 and k_2 , gives a satisfactory fit within measurement accuracy, although it might be argued that the fit at the end of the time interval is a bit poor (see Figure 6.16). Much effort was put in an investigation to improve the modelling of this tail, however, without success. Later, experimentalists reported that it is nearly impossible to keep experimental conditions constant, and – according to their judgement – the ‘poor’ fit towards the end of the time interval is likely to be a consequence of these varying experimental conditions. This means that no effort should be put in improving the fit at the end of the time interval as long as additional data are not available from experiments with constant experimental conditions.

In Table 6.12 the final estimates of the parameters are presented. For the sum of squared discrepancies (cf. (1.4)) all weights are set equal to 1.0, due to the assumed absolute measurement error. In the third and fourth column of this table the sizes of the independent and dependent confidence regions (cf. (1.25) and (1.26), with $\alpha = 0.05$) are reported. These sizes show that not all the parameters can be estimated with the desired accuracy. By desired we mean the accuracy the experimentalists wanted for reliable simulations. This means that additional measurements should be performed in order to obtain more accurate estimates. At the same time the sizes of the independent confidence regions are all smaller than the estimated parameters, which is something we rarely encountered in parameter estimation problems from real-life applications before additional measurements were performed (cf. the other case studies in this chapter). Apart from that, we were able to reject a couple of alternative models with the available measurements. For instance, at the beginning it was not known whether (6.66) should be considered as a fast equilibrium or as a reaction of the form:



However, it appears that this alternative leads to a better fit, the improvement was not sufficient to reject the null-hypothesis (cf. Appendix 1.C). Therefore, the model with the fast equilibrium assumptions was to be preferred. We will omit the F-ratio test here.

Parameters	Value	Ind.conf.reg.	Dep.conf.reg.
k_3	1.221×10^1	8.271×10^0	3.961×10^{-1}
k_4	8.616×10^{-2}	2.071×10^{-2}	2.708×10^{-3}
K	1.022×10^{-1}	4.511×10^{-2}	1.911×10^{-2}
klA	1.063×10^0	3.532×10^{-1}	1.730×10^{-1}
$y_6(0)$	3.599×10^{-1}	1.924×10^{-3}	1.401×10^{-4}

Table 6.12: Optimal parameters for ZLA-kinetics plus statistics after setting k_1 and k_2 equal to 0. The least squares estimate of $y_6(0)$ with more digits reads: 0.359903, which is close to the corresponding upperbound.

For this case study we also computed the matrix $Z^{(2)}$ from (5.1) in order to have a closer look into the parameter - state variable dependencies, the matrix is given by:

$$Z^{(2)} = \begin{pmatrix} 0.028 & 0.033 & 0.007 & 0.002 & 2.745 \\ 0.073 & 0.076 & 0.013 & 0.062 & 6.074 \\ 0.999 & 1.123 & 0.236 & 0.066 & 81.657 \\ 0.432 & 0.272 & 0.045 & 0.014 & 23.281 \\ 1.151 & 1.224 & 0.483 & 0.066 & 97.963 \\ 0.413 & 0.255 & 0.042 & 0.013 & 35.979 \\ 0.291 & 0.301 & 0.051 & 0.045 & 24.234 \end{pmatrix},$$

where the columns correspond to $k_1, k_2, k_3, k_4, K, klA$ and $y_6(0)$, and the rows to y_1, \dots, y_7 . The biggest entries are present in the last column, which means that a small relative change of the parameter $y_6(0)$ will affect the model responses of y_3 - y_7 considerably. The last row corresponds to the measured component, y_7 , its entries are in accordance with the results of Table 6.12. In order to decrease the confidence regions of k_3 and k_4 additional measurements for y_3 and y_5 should be performed. Such a design will also improve the reliability of K , but to a smaller extend. A design which focuses on the parameter with the relative biggest dependent confidence region, klA , does not exist.

6.4.4 Conclusions and remarks for further research

In the sections on ZLA-kinetics we managed to derive a model which fits the data and matches the available knowledge on this subject in this area of research. On the basis of the available data we were able to reject various, more complex models. The final model is compact and gives insight into the physically relevant aspects of the process.

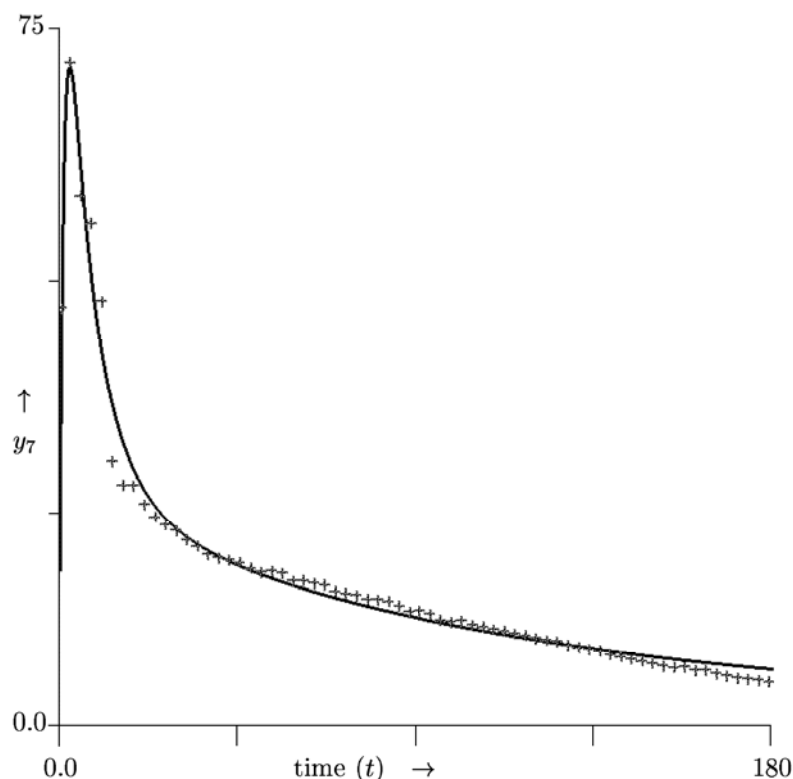


Figure 6.16: The data and the model responses for the carbon dioxide inflow, $y_7(t, \theta)$, are plotted for the optimal parameter values of Table 6.12.

In Section 6.4.3 we already saw that the results of parameter estimation and statistical analyses did not lead to accurate estimates for all parameters. This is due to the fact that the measurements only concern a single component of the state vector ($y_7(t)$). We showed which state variables should be measured during additional experiments to reduce the confidence regions. Additional measurements can also lead to a more complex model to be used in order to obtain a better insight into the investigated process.

Acknowledgement:

For a week the author was a guest at Akzo Nobel and started to explore the area between chemical processes and mathematical modelling. The cooperation and numerous discussions with Piet den Decker and Rein van der Hout –both at Akzo Nobel– were pleasant and constructive.

6.5 Water penetration in an aramide yarn[†]

In this section we study the mathematical modelling of water penetration in an *aramide yarn*. A model with only diffusion is compared with two models containing additional terms due to proposed chemistry during the absorption. The chemistry part models the binding of water in the yarn by either first and second order reaction kinetics, or a fast equilibrium. As a result we have three models with unknown parameters related to the diffusion, the reaction rates and the initial conditions of the yarn. The parameters are fitted to measurements of the weight of the yarn during the water absorption. Two of the three models could be rejected while the remaining model could be simplified without fitting significantly worse.

6.5.1 Introduction to the problem

We consider a long, thin and cylindrical yarn with length L , which is, after a long stay in a drying oven, put on a precision balance in a conditioned humid room. Here the yarn absorbs water from the surrounding air, it causes an increase of the yarn's weight, which is measured frequently during the absorption. After about 15 hours when the absorption is marginal the measurements stop.

The underlying physical process of the water absorption can be seen as a combination of diffusion and a reaction mechanism. The latter one describes a mechanism of water bound in open places inside the yarn.

6.5.2 Proposed models

Because the yarn is homogeneous we consider a section of the yarn with unit length. The model describing diffusion and a reaction mechanism for the binding of the water reads:

$$\frac{\partial u}{\partial t} = \nabla \cdot (D \nabla u) - k_1 u(c_a - w) + k_2 w, \quad (\text{on } \Omega \times]0, t_{end}]) \quad (6.84)$$

$$\frac{\partial w}{\partial t} = k_1 u(c_a - w) - k_2 w, \quad (\text{on } \Omega \times]0, t_{end}]) \quad (6.85)$$

$$u(x, t) = \alpha u_0, \quad (\text{on } \partial\Omega \times]0, t_{end}]) \quad (6.86)$$

$$u(x, 0) = w(x, 0) = 0. \quad (\text{on } \Omega) \quad (6.87)$$

Due to cylindrical symmetry Ω can be taken as the disc from a cross section of the yarn with radius R . The symbols u and w denote the concentrations in $[g/l]$ of the free and bound water, respectively. The quantity c_a denotes the concentration of places in the yarn where water can be bound and is yarn specific. (From (6.85) and the initial condition $w(x, 0) = 0$, it can be seen that $w(x, t)$ will never exceed c_a .) The numbers k_i ($i \in \{1, 2\}$) correspond to the reaction rates and α is a proportionality constant. The diffusion coefficient, $D \in \mathbb{R}^1$, is assumed to be constant throughout the yarn and due

[†] Work carried out in cooperation with R. van der Hout (Akzo Nobel Research, Arnhem).

to the symmetry, we can rewrite (6.84) as a 1-dimensional PDE, where $r \in [0, R]$ is the spatial coordinate, as:

$$\frac{\partial u}{\partial t} = \frac{D}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) - k_1 u(c_a - w) + k_2 w, \quad (\text{on } [0, R] \times]0, t_{end}]) \quad (6.88)$$

$$u(R, t) = \alpha u_0. \quad (\text{on }]0, t_{end}]) \quad (6.89)$$

Besides the model as introduced in (6.84)-(6.87) we introduce a model which contains the assumption that the chemical reaction is much faster than the diffusion. This means that we can consider a *steady state* instead of (6.85):

$$k_1 u(c_a - w) - k_2 w = 0,$$

or

$$w = \frac{k_1 c_a u}{k_1 u + k_2} = \frac{c_a u}{u + K}, \quad (6.90)$$

with $K = k_2/k_1$. Substitution of (6.90) in the original problem formulation (6.84)-(6.87) leads to:

$$\frac{\partial}{\partial t} \left(u + \frac{c_a u}{u + K} \right) = \frac{D}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right), \quad (\text{on } [0, R] \times]0, t_{end}]) \quad (6.91)$$

$$u(R, t) = \alpha u_0, \quad (\text{on }]0, t_{end}]) \quad (6.92)$$

$$u(r, 0) = w(r, 0) = 0. \quad (\text{on } \Omega) \quad (6.93)$$

In the subsequent sections we will refer to (6.84)-(6.87) as the original model and (6.91)-(6.93) as the simplified model. In Section 6.5.4 the performances of both models will be compared.

The parameters to be estimated are D , c_a , α and, depending on the model, k_1 and k_2 , or K . The parameters have to be estimated by fitting the numerical solution of the model to the measurements. During the experiments the weight of the yarn has been measured very precisely. The weight of the ‘dry’ yarn (after it comes from the drying oven) is denoted by W_0 . The weight after the yarn has been in the conditioned wet room for t hours reads:

$$W(t) = W_0 + 2\pi L \int_0^R (u(r, t) + w(r, t)) r dr, \quad (6.94)$$

where L is the length of the yarn. In case the simplified model is taken, (6.90) has to be substituted in (6.94).

For the yarn the *titer* (this is the weight per 10 000 m), the density, ρ , and the initial weight, W_0 , of the yarn are given. From these known quantities the radius, R , and the length, L , can be computed easily. Although the initial weight is known quite accurately, it will be considered as an unknown parameter later on, because this quantity contains a measurement error. Of course, if the estimated initial weight differs seriously from the measured initial weight, special attention should be paid.

6.5.3 Numerical implementation

For the numerical solution of the partial differential equations (PDEs) we used a *method of lines* in combination with a BDF solver for the resulting set of ordinary differential equations (ODEs). For the space discretisation we used the conservative form, with grid refinement in space towards the border and the centre of the yarn. The number of grid points is taken in such a way that the change of the solution after subsequent refinements is less than 1%. To our experience 15 spatial grid points are sufficient. After deriving the set of ODEs, the problem can easily be put in the *spIds* format (see [EHS95]).

The weight of the yarn, (6.94), is computed by a *trapezoidal rule*. Subsequently the unknown parameters are estimated by a least squares fit of the model responses to the measurements. The use of least squares estimation techniques is based on the assumption that the measurement errors are independent and normally distributed. The independence of the measurement errors plays an important role in Section 6.5.4 during the process of model discrimination. Three models will be considered for the model discrimination and data for one yarn during one experiment are available.

6.5.4 Model discrimination and parameter estimation

As a starting point for the model it is interesting to consider a simple diffusion equation without chemistry and fast equilibria. If we take (6.84)-(6.87) and set $k_1 = k_2 = c_a = 0.0$, the result is an equation with only diffusion and the remaining unknown parameters are: D , α and W_0 . The optimal fit for such a model is given in Figure 6.17. From this poor fit we can immediately see that the effect of diffusion only is not enough to describe the physical phenomenon.

The next step in the model discrimination process is to check whether the assumption about the fast equilibrium as proposed in (6.90) is valid. The result after fitting (6.84)-(6.87) to the data is plotted in the left graph of Figure 6.18. The corresponding residual sum is equal to 9.888×10^{-5} . If we repeat this numerical exercise with (6.91)-(6.93), then we obtain a residual sum of 1.828×10^{-3} and an optimal fit as shown in the right graph of Figure 6.18. Testing the null-hypothesis $K = k_2/k_1$ (cf. (6.90)) by means of the F-ratio test from Appendix 1.C leads to a rejection, which means that the assumption of a fast equilibrium does not hold.

Inspection of the graphs of Figure 6.18 provides a clue to another test. If the proposed model comes close to the true model and the estimated parameters come close to the true parameters, then the discrepancies get close to the measurement errors. The measurement errors are assumed to be stochastically independent, a test on the number of sign changes of the residuals in the graph at the right-hand side of Figure 6.18 will also reject the assumption of the fast equilibrium.

The optimal parameters corresponding to the model of (6.84)-(6.87), and the corresponding independent and dependent confidence regions, $\Delta^I \theta$ and $\Delta^P \theta$, respectively (cf. Section 1.6) are given in Table 6.13. From this table we may conclude that only W_0 and c_a can be estimated with a reasonable accuracy; for all other parameters $\Delta^I \theta_i$ is at least

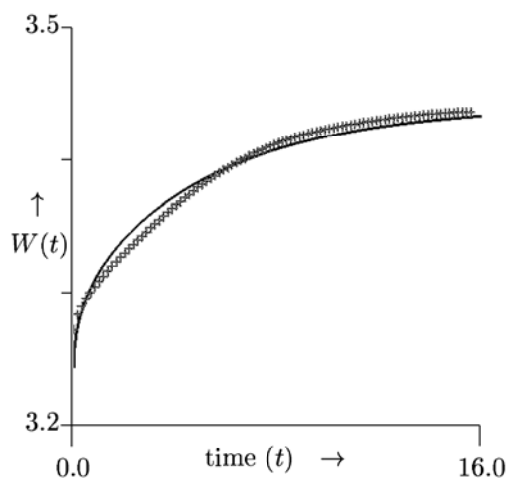


Figure 6.17: Optimal fit in case the physical process is modelled with only diffusion, i.e., (6.84)-(6.87), with $k_1 = k_2 = c_a = 0.0$.

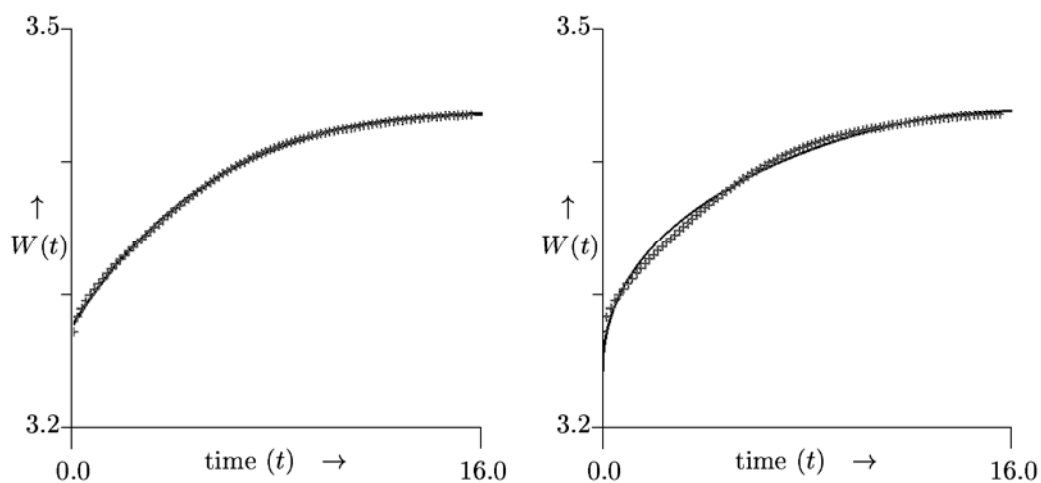


Figure 6.18: Optimal fit in case of general chemistry (6.84)-(6.87) (left) and the result with the fast equilibrium assumption (6.90) (right).

one order of magnitude bigger than the corresponding estimated value. The result that only 2 out of 6 unknown parameters can be estimated within reasonable accuracy seems a bit of a disappointment, in particular if we know that the parameter W_0 was already known quite accurate before the parameter estimation started. However, things turn

θ	$\hat{\theta}$	$\Delta' \theta$	$\Delta'' \theta$
D	1.401×10^{-7}	2.760×10^{-6}	2.573×10^{-9}
α	5.906×10^1	1.176×10^3	5.060×10^{-1}
W_0	3.275×10^0	6.546×10^{-3}	3.838×10^{-4}
c_a	6.857×10^{-2}	1.526×10^{-2}	2.678×10^{-4}
k_1	7.024×10^2	1.407×10^4	1.159×10^1
k_2	9.091×10^{-11}	5.618×10^{-2}	1.627×10^{-3}
$S(\hat{\theta})$	9.888×10^{-5}		

Table 6.13: Optimal parameter values and corresponding statistics for the parameters of model (6.84)-(6.87), which correspond to the left graph of Figure 6.18.

better if we become aware that by the computation we were able to distinguish between models and that we are now able to indicate how much information can be retrieved from the measurements.

When we consider the value of k_2 in Table 6.13, we see that this value is close to zero. This means, according to the model equations, that once free water is bound, the reverse process is very slow. We can go further and check whether the inverse process is essential. To be more precise, we want to test the null hypothesis $H_0 : k_2 = 0.0$ versus the alternative $H_1 : k_2 > 0.0$. With the theory from Appendix 1.C and the residual sums of squared residuals from the Tables 6.13 and 6.14, we can derive easily that the null hypothesis H_0 is not rejected ($N = 94$).

θ	$\hat{\theta}$	$\Delta' \theta$	$\Delta'' \theta$
D	1.401×10^{-7}	2.482×10^{-6}	2.388×10^{-9}
α	5.906×10^1	1.061×10^3	4.710×10^{-1}
W_0	3.275×10^0	5.956×10^{-3}	3.583×10^{-4}
c_a	6.857×10^{-2}	7.519×10^{-3}	2.503×10^{-4}
k_1	7.024×10^2	1.275×10^4	1.079×10^1
$S(\hat{\theta})$	9.897×10^{-5}		

Table 6.14: Optimal parameter values and corresponding statistics for the parameters of model (6.84)-(6.87) after fixing k_2 to zero.

6.5.5 Conclusions

From this case study it turned out that water penetration in an aramide yarn cannot be described by diffusion only. Further, if we add a reaction mechanism to describe the binding of water inside the yarn, then this process cannot be considered as a fast

equilibrium. The reaction part which describes loosening of the water is not essential in the process. Finally, 5 parameters remain to model the penetration. These parameters cannot be estimated within an acceptable accuracy on the basis of the available data. Additional measurement from other similar yarns will hardly give any additional information, because –except for k_1 – all the other parameters depend on the structure of the yarn. More accurate results are expected from experiments with different humidities in the conditioned room, u_0 , and yarns with the same structure but different radii.

6.6 Macroeconomic time series

6.6.1 Introduction

This section concerns a case study from econometrics. We consider Indian macroeconomic time series of currency notes in circulation between January 1970 and March 1985. These monthly data are taken from [Ray88]. In econometric sciences such series are studied by means of linear autoregressive moving average (*ARMA*) and nonlinear methods, e.g. self-exciting threshold autoregressive (*SETAR*) methods. For an introduction on such methods see e.g. [BD87]. *ARMA* and *SETAR* methods are applied to the above mentioned currency data in [BG97], in order to model the macroeconomic process and to retrieve reliable predictions. In this section we use another approach to account for the data by deriving a family of candidate models, which are continuous and given by algebraic equations only. These models are fitted to the data in a least squares sense. The best fitting model is used for prediction purposes over the period April 1985 till March 1986. These predictions are compared with the real data from that period and also with the predictions given in [BG97].

6.6.2 Derivation of candidate models

Direct inspection of the data (see Figure 6.19) gives already an indication for the type of models we want to endeavour. The data show the presence of an exponential growth with a periodic perturbation due to seasonal effects. Three straightforward ways to combine the periodic behaviour with the exponential growth are an additive, multiplicative or exponential relationship, as given in the following formulae:

$$y(t) = a + f \sin(d t + e) + b \exp(c t) , \quad (6.95)$$

$$y(t) = a + b(f + \sin(d t + e)) \exp(c t) , \quad (6.96)$$

$$y(t) = a + b \exp(c t(f + \sin(d t + e))) . \quad (6.97)$$

Here t denotes the time (in months), y is the amount of currency notes in circulation and a to f are parameters which are estimated by fitting the model to the given data. From the sum of squares for the best fit of each model, as given in Table 6.15, it can be derived by an F-ratio test that the additive model fits significantly poorer ($\alpha < 0.05$) than the other two. From the same test it cannot be decided whether the multiplicative model performs significantly better than the exponential model.

A closer look at the data of Figure 6.19 shows that the periodic component is not symmetric, but tends towards a *saw-tooth* shape. The saw-tooth can mathematically be expressed by the series:

$$\text{st}(x) = \sum_{i=1}^{\infty} (-1)^{(i+1)} \frac{\sin(i x)}{i} . \quad (6.98)$$

Due to the decrease which is steeper than the increase in every periodic cycle we replace the periodic term in any of the models (6.95)–(6.97) by a finite subsequence of (6.98) with

Model	Sum of squares
Additive (6.95)	3.083×10^7
Multiplicative (6.96)	1.967×10^7
Exponential (6.97)	2.122×10^7

Table 6.15: Sum of squared discrepancies for the models (6.95)-(6.97) .

N_{st} terms. Due to the freedom with respect to N_{st} , we obtain many models which can be tested and compared. We fit the additive, multiplicative and exponential models for different numbers of N_{st} . Because the fits become poor for $N_{st} > 6$, we stop increasing N_{st} further.

From the investigated models, the model which gave smallest sum of squared discrepancies reads:

$$y(t) = a + b \left(f + \sum_{i=1}^4 -1^{(i+1)} \frac{\sin(d t + e)}{i} \right) \exp(c t) . \quad (6.99)$$

As before, the additive model performed significantly poorer, but the multiplicative model with the smallest least squares did not fit significantly better than multiplicative models with $N_{st} \leq 5$, or the exponential model with $N_{st} \leq 4$. Another approach to choose N_{st} is to consider it as an integer parameter and change the problem into a mixed integer minimisation problem. However, the solution of such a problem is beyond the scope of the present thesis.

The estimated parameter values corresponding to model (6.99) are given in Table 6.16, the corresponding model responses are given by the solid line in Figure 6.19 and the corresponding sum of squares equals: 1.788×10^7 .

Parameters	Value	Ind.conf.reg.	Dep.conf.reg.
a	2.127×10^3	4.278×10^2	8.437×10^1
b	6.184×10^1	2.032×10^1	5.523×10^{-1}
c	1.359×10^{-2}	7.662×10^{-4}	5.909×10^{-5}
d^a	5.256×10^{-1}	5.615×10^{-3}	1.250×10^{-3}
e	-3.971×10^{-1}	8.343×10^{-1}	1.859×10^{-1}
f	2.863×10^1	8.990×10^0	2.556×10^{-1}

Table 6.16: Optimal parameters for model (6.99) fitted to measured currency notes in circulation.

^a Note that $2\pi/12 = 0.523599 \dots$

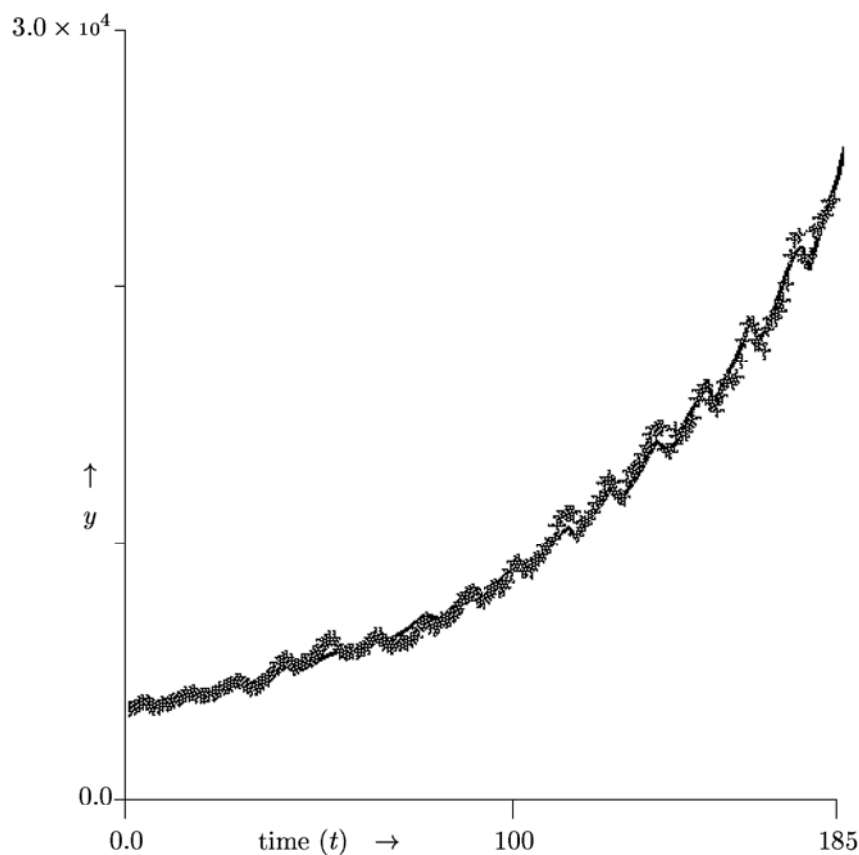


Figure 6.19: The data (indicated by '+') and the model responses from Eq. (6.99) for the currency notes in circulation for the optimal parameter values of Table 6.16.

6.6.3 Comparison of prediction results

One of the targets of fitting the models or using ARMA or SETAR methods is their use for prediction purposes. On the basis of the monthly measurements (January 1970 till April 1985) we predict the currency notes for the period May 1985 till March 1986. For the ARMA and SETAR methods this is done and described in [BG97], we compare their results with predictions based on (6.99) and the parameters of Table 6.16. The results of these three methods are given in Table 6.17.

The differences between the second and third column of Table 6.17 have been discussed in [BG97], where the authors stress that the SETAR outcome is sensitive to the

Month	Obs. values	ARMA discr.	SETAR discr.	(6.99) discr.
April 1985	24129	86.0	129.7	-211.5
May	24852	47.0	194.6	-366.2
June	24775	258.9	514.7	385.1
July	23755	542.2	619.3	656.8
August	23357	608.7	827.9	114.0
September	23036	722.1	1183.0	983.9
October	23783	1035.6	1463.0	908.1
November	24486	647.7	1124.7	518.6
December	24454	900.8	1314.3	1242.1
January 1986	24364	1227.0	1613.3	1827.1
February	24823	1237.1	1574.6	1839.0
March	25519	1132.6	1228.8	1887.9
Sum of sq. discr.		7.885×10^6	1.451×10^7	1.466×10^7

Table 6.17: Observed values, and the discrepancies between the predicted and the observed values with three different prediction methods for the period April 1985-March 1986.

choice of the delay and the threshold value in this model. This means that their results are not found in a straightforward manner, but the method contains additional freedom which was used to improve the fits. The predictions of the SETAR model are better than the results of (6.99), but the difference is marginal and far from significant.

In this comparison it should be added that the best fitting SETAR model uses 12 parameters compared to 6 for the multiplicative model. On the other hand, the SETAR models can be used in a wider context than the model we deduced from inspecting the data. The method we followed is easy to use and as shown here competitive in this case study from literature, although it might be less appropriate in situations with more complex data sets.

6.6.4 Concluding remarks

In this case study we compared ARMA and SETAR methods with a simpler model we derived from the data by inspecting them. The model as described here is competitive with the SETAR method in predicting currency notes, while at the same time it needs less parameters and is easier to use. Our derivation of the model requires some experience and insight in the data and the result is a bit of a special purpose model, which cannot be used for an arbitrary time series. The effort to derive a model as we did is worth considering if many predictions have to be made or the process is studied thoroughly. For an exploratory study ARMA or SETAR models are favourite.

6.7 The DOW problem

6.7.1 Introduction

This problem was originally formulated by the Dow Chemical Company and studied already by various researchers in the 80's. A number of solutions to this problem from five different research groups is presented and compared in [BDB86].

The model is described by a system of 10 stiff differential-algebraic equations (DAEs) and contains 9 unknown parameters to be estimated. Besides the model we have three sets of data from a batch reactor. The sets differ by initial concentrations and temperatures. The error structure of the data is not a priori known. We assume that the measurement errors are normally distributed. Further, it is likely that, due to the way the measurements were constructed, their errors are correlated. Measurement of four chemical compounds are available.

The purpose here is to use the same problem as a case study for our parameter estimation approach. First, we compute the estimates taking the 4×4 covariance matrix V (cf. (3.11)) (i) equal to the identity matrix, (ii) diagonal with unknown entries (Section 3.3.2) and (iii) a full matrix with unknown entries (Section 3.5). Second, we compare our results with those reported in [BDB86]. Third, we compare the parameter estimates obtained when taking a full or a diagonal covariance matrix. Finally, we drop the model assumptions as they are in the original formulation, formulate a more general model and compare the corresponding results with those from the original model.

6.7.2 Description of the problem

For the formulation of the problem we refer to [BDB86] and restrict ourselves to giving the set of DAEs. Although many model assumptions are present, we did not look into the validity of these assumptions. The reader interested in the chemical background of the equations and the related assumptions is referred to the original article ².

The model is mathematically expressed by 6 differential and 4 algebraic equations. As in Chapter 1, the state variables are denoted by y and depend on time and the unknown parameters.

$$\frac{dy_1}{dt} = -k_2 y_2 y_8 \quad (6.100)$$

$$\frac{dy_2}{dt} = -k_1 y_2 y_6 + k_{-1} y_{10} - k_2 y_2 y_8 \quad (6.101)$$

$$\frac{dy_3}{dt} = k_2 y_2 y_8 + k_1 y_4 y_6 - \frac{1}{2} k_{-1} y_9 \quad (6.102)$$

$$\frac{dy_4}{dt} = -k_1 y_4 y_6 + \frac{1}{2} k_{-1} y_9 \quad (6.103)$$

²The cited article contains two mistakes in the DAEs on page 31: there, the first minus sign in the right-hand of both equation (3) and (5) should be omitted.

$$\frac{dy_5}{dt} = k_1 y_2 y_6 - k_{-1} y_{10} \quad (6.104)$$

$$\frac{dy_6}{dt} = -k_1 y_6 [y_2 + y_4] + k_{-1} [y_{10} + \frac{1}{2} y_9] \quad (6.105)$$

$$y_7 = -[Q^+] + y_6 + y_8 + y_9 + y_{10} \quad (6.106)$$

$$y_8 = \frac{K_2 y_1}{K_2 + y_7} \quad (6.107)$$

$$y_9 = \frac{K_3 y_3}{K_3 + y_7} \quad (6.108)$$

$$y_{10} = \frac{K_1 y_5}{K_1 + y_7} \quad (6.109)$$

The unknown parameters k_{-1} , k_1 and k_2 are assumed to be temperature dependent via Arrhenius' law, which yields three extra parameters:

$$k_i = \alpha_i \exp(-E_i/(RT)) , \quad i \in \{-1, 1, 2\} . \quad (6.110)$$

Here α_i , given in $[mol/(kg*hrs)]$, is the pre-exponential factor and E_i , $[cal/mol]$ is the activation energy.

Summarising, we have nine parameters, given by the vector $\theta = [\alpha_1, E_1, \alpha_2, E_2, \alpha_{-1}, E_{-1}, K_1, K_2, K_3]$. The vector of state variables corresponding to the chemical compounds under consideration is $y = [HA, BM, HABM, AB, MBMH, M^-, H^+, A^-, ABM^-, MBM^-]$. The quantity $[Q^+]$ in equation (6.106) is a concentration which is assumed to be constant during the reactions.

The initial conditions read:

$$\begin{aligned} y_5(0) &= 0.0 , \\ y_6(0) &= [Q^+] = 0.00131 , \\ y_7(0) &= \frac{1}{2} \left(-K_2 + \sqrt{K_2^2 + 4K_2 y_1(0)} \right) , \\ y_8(0) &= y_7(0) , \\ y_9(0) &= 0.0 , \\ y_{10}(0) &= 0.0 . \end{aligned}$$

The initial condition $y_9(0)$ is not consistent with the initial data of the second and third experiment, but the DAE solver will correct this immediately. The initial guess for the unknown parameters, θ_{ini} , provided by the Dow description, reads:

$$\begin{aligned} \alpha_1 &= 2.0E13 , \\ E_1 &= 2.0E3 , \\ \alpha_2 &= 2.0E13 , \\ E_2 &= 2.0E3 , \end{aligned}$$

$$\begin{aligned}
\alpha_{-1} &= 4.3E15, \\
E_{-1} &= 2.0E3, \\
K_1 &= 1.0E-17, \\
K_2 &= 1.0E-11, \\
K_3 &= 1.0E-17,
\end{aligned}$$

where K_i , $i \in \{1, 2, 3\}$ is given in: $[mol/kg]$.

As in Section 6.1 we take the logarithm of the pre-exponential factor, divide the activation energies by 1000 and use the reparametrisation from (6.16). The values of K_i ($i \in \{1, 2, 3\}$) also lead to the decision to replace these parameters by their logarithms. This completes the model specification of the parameter estimation problem.

6.7.3 Data

The available data originate from three different experiments. During each experiment the temperature is constant, while from experiment to experiment the temperatures vary (40, 67 and 100°C). During the experiment, data from four different components are observed; three species are measured and the measured values are adjusted according to a conservation law. The value of the fourth data point is derived by making use of an additional relation. This history with respect to the origin of the data make it likely that the ‘measurements’ are correlated. Therefore, we will estimate –apart from the 9 unknown parameters– a 4×4 covariance matrix as explained in Section 3.5.

From the estimated covariance matrix we will investigate the dependence of the measurement errors. We will also compare the results obtained by a full covariance matrix, with those for a diagonal covariance matrix.

6.7.4 Results

First case: V is the identity matrix

By making this choice for the covariance matrix, we neglect the information about the history of the data. Nevertheless we start this way as a first exploration of the model. The OLS estimates and the corresponding confidence regions are given in Table 6.18, the optimal fits are shown in Figure 6.20.

In the comparison of Biegler et al. [BDB86], only one research group used ordinary least squares, but they considered 3 measured components, 8 parameters and 3 ODEs. Their residual sum equals 0.233, which is not significantly different from the result reported in Table 6.18. The estimated parameters and the corresponding fits of this group are in close correspondence with the results given here. However, this group did not give the corresponding confidence regions.

The confidence regions in Table 6.18 are satisfactory, except for $\ln(\tilde{\alpha}_{-1})$ and $\ln(K_i)$ ($i = 1, 2, 3$). Taking into account the equations for y_8 , y_9 and y_{10} ((6.107)-(6.109)) and

	final est. ($\hat{\theta}$)	independent confidence regions ($\Delta^I \theta$)	dependent confidence regions ($\Delta^D \theta$)
$\ln(\tilde{\alpha}_1)$	-0.8134	3.9371×10^{-1}	5.7949×10^{-2}
\tilde{E}_1	18.5632	2.0815×10^0	4.8385×10^{-1}
$\ln(\tilde{\alpha}_2)$	-1.1331	1.6464×10^{-1}	3.1420×10^{-2}
\tilde{E}_2	18.8592	6.2497×10^{-1}	2.7807×10^{-1}
$\ln(\tilde{\alpha}_{-1})$	-12.0012	1.6649×10^3	1.1487×10^{-1}
\tilde{E}_{-1}	26.1344	1.7712×10^0	1.0214×10^0
$\ln(K_1)$	-37.2129	2.1014×10^7	3.8612×10^{-2}
$\ln(K_2)$	-28.0053	2.1015×10^7	1.1486×10^{-1}
$\ln(K_3)$	-37.5920	2.1014×10^7	3.9349×10^{-2}
$S(\theta)$	0.3958		

Table 6.18: Final estimates of θ for the case $V = I_4$ plus confidence regions (cf. (1.25) and (1.26), with $\alpha = 0.05$).

the size of K_i ($i = 1, 2, 3$), it is no surprise that the parameters $\ln(K_i)$ ($i = 1, 2, 3$) cannot be estimated accurately.

Our confidence regions can be compared with the confidence regions reported for other choices of V as found below.

Second case: V is diagonal and unknown

Here we estimate both the weights and the parameters as described in Section 3.3.2. The most likely weights, $1/\hat{\sigma}_j$ in (3.22), are:

$$w = (31.4258, 23.3746, 28.8777, 41.1157)^T.$$

We see immediately that the weights are all of the same order of magnitude, which means that no big changes with respect to the optimal parameters and the corresponding fits are expected. The parameters and the confidence regions are given in Table 6.19. The graphs of these fits are not shown because they are almost identical to those in Figure 6.20.

Instead of (3.19) we consider:

$$\tilde{\mathcal{L}}(\theta) = \prod_{j=1}^q \left(\frac{1}{r_j} \sum_{i=1}^r D_{ij}^2 \right)^{\frac{r_j}{2}}, \quad (6.111)$$

which makes no difference for the minimisation, but the result can be interpreted geometrically: the outcome of (6.111) equals the volume of a box in the data space, whose edges equal the estimated standard deviation of the corresponding measurement error. The (natural) logarithm of this volume, where $\sum_{j=1}^4 r_j = 339$, is -1157.2.

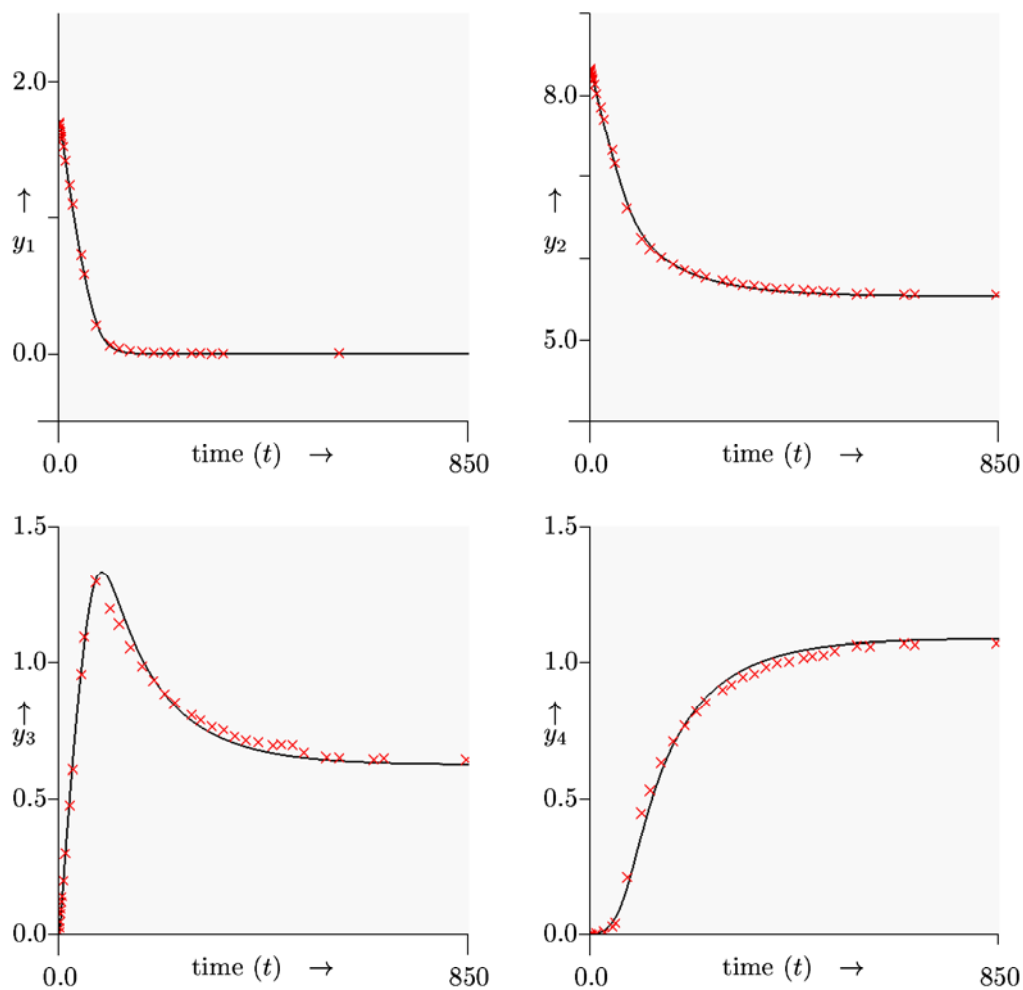


Figure 6.20: Optimal fits for y_1 to y_4 during the first experiment at 40°C for the case $V = I_4$.

In [BDB86] three research groups considered this estimation problem with an unknown diagonal covariance matrix with four regressed components. However, these groups took different values for the heteroscedasticity. The heteroscedasticity, denoted by $\gamma_i \in [0, 2]$, is a measure for the influence of the relative and the absolute error. It is 0 in the case of absolute and 2 in the case of relative measurement errors. One research group in [BDB86] also estimated the heteroscedasticities. The two other groups chose γ_i ($i = 1, \dots, 4$) equal to 0 or 1, respectively. The calculations in this section are performed

	final est. ($\hat{\theta}$)	independent confidence regions ($\Delta^I \theta$)	dependent confidence regions ($\Delta^D \theta$)
$\ln(\tilde{\alpha}_1)$	-0.8048	3.9362×10^{-1}	5.4678×10^{-2}
\tilde{E}_1	18.6136	2.0027×10^0	4.5443×10^{-1}
$\ln(\tilde{\alpha}_2)$	-1.1409	1.7147×10^{-1}	3.3010×10^{-2}
\tilde{E}_2	18.8057	6.5829×10^{-1}	2.9119×10^{-1}
$\ln(\tilde{\alpha}_{-1})$	-22.4686	6.1303×10^7	1.1671×10^{-1}
\tilde{E}_{-1}	25.7828	1.6628×10^0	1.0453×10^0
$\ln(K_1)$	-47.2453	5.7088×10^7	3.5545×10^{-2}
$\ln(K_2)$	-27.6020	1.4595×10^7	1.1671×10^{-1}
$\ln(K_3)$	-47.6189	5.7089×10^7	3.5957×10^{-2}

Table 6.19: Final estimates of θ for the case V is diagonal and unknown, plus confidence regions (cf. (1.25) and (1.26), with $\alpha = 0.05$).

assuming absolute measurement errors. Comparison of our results with the results with zero heteroscedasticities showed a close correspondence with respect to the fits and the parameters. We encountered different values for the computed confidence regions. However, the paper does not give an explanation how the reported values were obtained. Also the results obtained by the group which took all heteroscedasticities equal to one are close to our results. The confidence regions by this group are also in accordance with our results

Third case: V is full and unknown

Here we follow the approach from Section 3.6, estimating the parameters by minimising the determinant of the moment matrix (cf. (3.10)). It should be noted that in this case the measured components are not the same for all samples, $N < qr$. Therefore, we first compute the moment matrix, and then we divide the entry M_{ij} by the number of samples containing measurements related to both D_{li} and D_{lj} ($l = 1, \dots, r$). The resulting matrix is a biased estimator of the covariance matrix. The bias has no influence on the parameter estimates. Finally, we minimise the square root of the determinant of this biased covariance matrix, again after correcting for the non-constant samples.

The biased estimate of the covariance matrix reads:

$$\hat{V} = \begin{bmatrix} 1.0556 \times 10^{-03} & 1.2917 \times 10^{-03} & -8.6404 \times 10^{-04} & -2.0152 \times 10^{-04} \\ 1.2917 \times 10^{-03} & 1.9221 \times 10^{-03} & -4.0993 \times 10^{-04} & -7.5342 \times 10^{-04} \\ -8.6404 \times 10^{-04} & -4.0993 \times 10^{-04} & 1.1618 \times 10^{-03} & -4.1305 \times 10^{-04} \\ -2.0152 \times 10^{-04} & -7.5342 \times 10^{-04} & -4.1305 \times 10^{-04} & 5.7076 \times 10^{-04} \end{bmatrix}$$

This matrix shows that independence of the measurement errors is unlikely, due to the fact that the matrix is not even diagonal dominant.

The optimal parameters and the corresponding confidence regions are given in Table 6.20, the corresponding fits do not differ significantly from the plots given in Figure 6.20.

	final est. ($\hat{\theta}$)	independent confidence regions ($\Delta^I \theta$)	dependent confidence regions ($\Delta^D \theta$)
$\ln(\tilde{\alpha}_1)$	-0.7892	2.5698×10^{-1}	5.5065×10^{-2}
\tilde{E}_1	18.6555	1.5277×10^0	4.5556×10^{-1}
$\ln(\tilde{\alpha}_2)$	-1.1509	1.5588×10^{-1}	3.8189×10^{-2}
\tilde{E}_2	18.7500	7.5025×10^{-1}	3.3533×10^{-1}
$\ln(\tilde{\alpha}_{-1})$	-22.2395	3.3500×10^7	1.2862×10^{-1}
\tilde{E}_{-1}	25.4406	1.6515×10^0	1.1621×10^0
$\ln(K_1)$	-47.2155	3.6002×10^7	3.5108×10^{-2}
$\ln(K_2)$	-27.8259	1.5265×10^7	1.2862×10^{-1}
$\ln(K_3)$	-47.5814	3.6002×10^7	3.5237×10^{-2}

Table 6.20: Final estimates of θ for the case V is full, symmetric and unknown, plus confidence regions (cf. (1.25) and (1.26), with $\alpha = 0.05$).

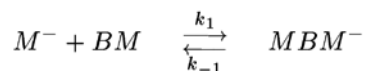
The (natural) logarithm of the volume of the corresponding box in the dataspace (cf. (6.111)) equals -1361.8, which is an improvement –as expected– of the result with a diagonal covariance matrix. However, the estimated parameters and their confidence regions do not change considerably from one case to the other. We can state that the estimates and the corresponding confidence regions are not sensitive to the three choices of V we made.

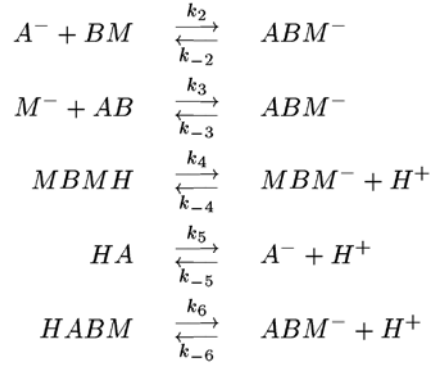
In Biegler et al. [BDB86] only one group performs the regression with a full covariance matrix. However, they only consider two regressed components, which is an oversimplification of the problem and makes comparison irrelevant.

6.7.5 More general model equations

In the article by Biegler et al. ([BDB86]), many assumptions –with respect to the rapid acid-base reactions, $k_{-2} = 0$, $k_3 = k_1$ and $k_{-3} = 1/2k_{-1}$ – are already made. The authors motivate this by stating that a more general model would have too many parameters that cannot be estimated. This is true, but to our opinion it is a more general approach to start with a model which contains less assumptions and more parameters. Such an approach should be a starting point for a step by step process of checking assumptions, eliminating parameters and making model simplifications.

A schematic representation of the reactions without the additional assumptions reads:





The mathematical model corresponding to the above scheme yields:

$$\frac{d[HA]}{dt} = -k_5[HA] + k_{-5}[A^-][H^+] \quad (6.112)$$

$$\begin{aligned} \frac{d[BM]}{dt} = & -k_1[M^-][BM] + k_{-1}[MBM^-] - \\ & k_2[A^-][BM] + k_{-2}[ABM^-] \end{aligned} \quad (6.113)$$

$$\frac{d[HABM]}{dt} = -k_6[HABM] + k_{-6}[ABM^-][H^+] \quad (6.114)$$

$$\frac{d[AB]}{dt} = -k_3[M^-][AB] + k_{-3}[ABM^-] \quad (6.115)$$

$$\frac{d[MBMH]}{dt} = -k_4[MBMH] + k_{-4}[MBM^-][H^+] \quad (6.116)$$

$$\begin{aligned} \frac{d[M^-]}{dt} = & -k_1[M^-][BM] + k_{-1}[MBM^-] - \\ & k_3[M^-][AB] + k_{-3}[ABM^-] \end{aligned} \quad (6.117)$$

$$\begin{aligned} \frac{d[H^+]}{dt} = & k_5[HA] - k_{-5}[A^-][H^+] + \\ & k_4[MBMH] - k_{-4}[MBM^-][H^+] + \\ & k_6[HABM] - k_{-6}[ABM^-][H^+] \end{aligned} \quad (6.118)$$

$$\begin{aligned} \frac{d[A^-]}{dt} = & -k_2[A^-][BM] + k_{-2}[ABM^-] + \\ & k_5[HA] - k_{-5}[A^-][H^+] \end{aligned} \quad (6.119)$$

$$\begin{aligned} \frac{d[ABM^-]}{dt} = & k_2[A^-][BM] - k_{-2}[ABM^-] + \\ & k_3[M^-][AB] - k_{-3}[ABM^-] + \\ & k_6[HABM] - k_{-6}[ABM^-][H^+] \end{aligned} \quad (6.120)$$

$$\begin{aligned} \frac{d[MBM^-]}{dt} = & k_1[M^-][BM] - k_{-1}[MBM^-] + \\ & k_4[MBMH] - k_{-4}[MBM^-][H^+] \end{aligned} \quad (6.121)$$

For the all reaction rates k_i ($i = \pm 1, \dots, \pm 6$) we have an Arrhenius' relation as in (6.110). Further, we reparametrised the pre-exponential factor as in (6.16). The results of the regression with this more general model for the case $V = I_4$ are given in Table 6.22 and Figure 6.21. The residual sum equals 0.2352 (with $N - m = 339 - 24$ degrees of freedom) which is significantly better than 0.3958 (with $N - m = 339 - 9$) from Table 6.18, because the corresponding ratio:

$$X = \frac{(0.3958 - 0.2352)/(24 - 9)}{0.2352/(339 - 24)} = 14.34 .$$

and the corresponding upper quantile, $\mathcal{F}_{0.05}(15, 315)$, equals 1.70.

We will also perform the F-ratio test as introduced at the end of Appendix 1.C. The residual sums after the data have been split are listed in Table 6.21. As in Appendix 1.C

	$N_1 = 167$	$N_2 = 172$
$m_1 = 24$	0.06695	0.1561
$m_2 = 9$	0.1223	0.2682

Table 6.21: Residual sums for the models (6.112)- (6.121) with 24 parameters and (6.100)-(6.109) with 9 parameters when the data are split in two disjunct sets.

we compute the ratios:

$$X_{1,2} = \frac{0.06695/(167 - 24)}{0.2682/(172 - 9)} = 0.2845 ,$$

and

$$X_{2,1} = \frac{0.1561/(172 - 24)}{0.1223/(167 - 9)} = 1.3626 .$$

For the lowerbound corresponding to $X_{1,2}$ at a confidence level of 0.95, we get $1/\mathcal{F}_{0.0125}(163, 143) = 0.6929$ and the upperbound for $X_{1,2}$ reads $\mathcal{F}_{0.0125}(148, 158) = 1.4379$. These results (cf. (1.38)) are in accordance with the results we obtained from the other test.

6.7.6 Concluding remarks

In this section we solve the problem reported in [BDB86] in various ways and, where possible, we compare our results with the results in this article. With respect to the fitness criterion three different situations have been studied: the covariance matrix is (i) the identity matrix, (ii) diagonal with unknown entries and (iii) full with unknown

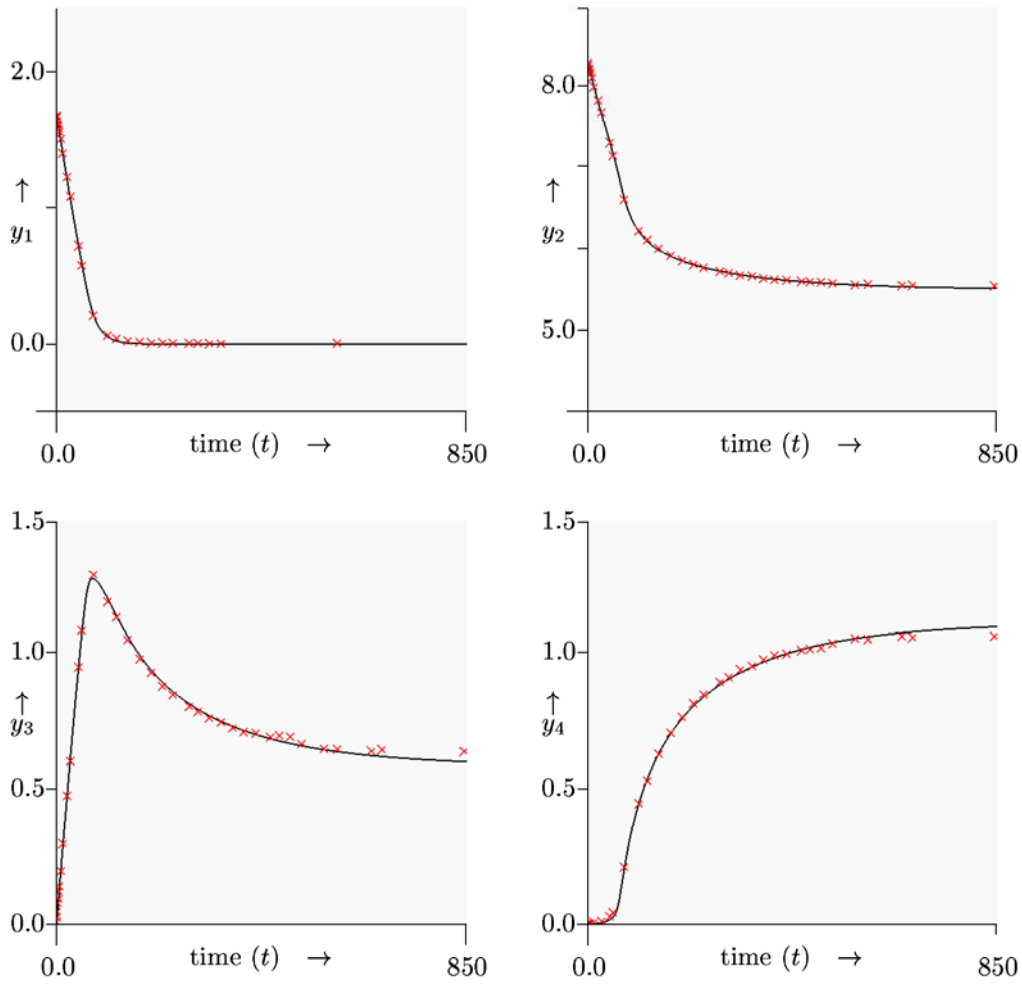


Figure 6.21: Experimental (\times) and numerical results obtained through (6.112)-(6.121), $V = I_4$ and the parameters of Table 6.22.

entries. The estimated entries for an unknown diagonal covariance matrix were all of the same order of magnitude. For a full covariance matrix we find that this matrix was not diagonal dominant and, thus, it is likely that the measurement errors are correlated, as was expected from the description how they were generated.

In general, our fits and the parameter results are in correspondence with the results of [BDB86]. The more general case of a full 4×4 -matrix (which takes dependence of

	final est. ($\hat{\theta}$)	independent confidence regions ($\Delta^I \theta$)	dependent confidence regions ($\Delta^D \theta$)
$\ln(\tilde{\alpha}_1)$	9.7790	2.0566×10^4	4.3305×10^{-2}
$\ln(\tilde{E}_1)$	36.5900	1.2971×10^5	3.7914×10^{-1}
$\ln(\tilde{\alpha}_2)$	1.0223	1.1435×10^{-1}	3.7432×10^{-2}
$\ln(\tilde{E}_2)$	18.9920	9.5607×10^{-1}	3.3461×10^{-1}
$\ln(\tilde{\alpha}_3)$	4.9937	2.5584×10^1	4.4374×10^{-2}
$\ln(\tilde{E}_3)$	16.3235	1.7288×10^2	3.8994×10^{-1}
$\ln(\tilde{\alpha}_{-1})$	17.9814	2.7403×10^4	4.3313×10^{-2}
$\ln(\tilde{E}_{-1})$	27.4450	1.4187×10^5	3.7924×10^{-1}
$\ln(\tilde{\alpha}_{-2})$	10.2130	2.3841×10^4	4.0345×10^{-1}
$\ln(\tilde{E}_{-2})$	28.1234	1.1722×10^5	3.3370×10^0
$\ln(\tilde{\alpha}_{-3})$	12.1521	2.3834×10^4	4.1852×10^{-2}
$\ln(\tilde{E}_{-3})$	12.5168	1.1720×10^5	3.6596×10^{-1}
$\ln(\tilde{\alpha}_4)$	-4.0824	9.9917×10^{-1}	6.1516×10^{-2}
$\ln(\tilde{E}_4)$	20.2311	4.9609×10^0	5.8733×10^{-1}
$\ln(\tilde{\alpha}_5)$	-1.1080	4.7799×10^{-1}	8.7616×10^{-2}
$\ln(\tilde{E}_5)$	14.0946	2.6856×10^0	7.9105×10^{-1}
$\ln(\tilde{\alpha}_6)$	7.9283	3.2564×10^4	4.3690×10^{-2}
$\ln(\tilde{E}_6)$	10.4588	2.7917×10^5	3.8273×10^{-1}
$\ln(\tilde{\alpha}_{-4})$	30.6320	1.3157×10^4	4.3311×10^{-2}
$\ln(\tilde{E}_{-4})$	4.6011	7.0823×10^4	3.7922×10^{-1}
$\ln(\tilde{\alpha}_{-5})$	20.7319	2.6032×10^3	5.7788×10^{-1}
$\ln(\tilde{E}_{-5})$	13.5023	2.3525×10^4	4.9884×10^0
$\ln(\tilde{\alpha}_{-6})$	42.5718	3.0832×10^4	4.3690×10^{-2}
$\ln(\tilde{E}_{-6})$	0.3731	3.2713×10^5	3.8273×10^{-1}
$S(\theta)$	0.2352		

Table 6.22: Final estimates of θ and their confidence regions for the model of (6.112)-(6.121). The corresponding fits are found in Figure 6.21

the measurement errors into account) was not dealt with in that article. However, our results with respect to the parameters and the fits led to marginal changes for the different choices of V .

When we drop the assumptions proposed in the original formulation of the problem and solve the regression problem with a general model, we get a fit which is significantly better. However, the problem is poorly conditioned.

Appendix 6.A

This appendix contains an example model file as it is needed for the parameter estimation program, spIds ([EHS95]). In fact, the model file shown was used for the resin problem of Section 6.1. The lines starting with an “#” are comment lines. Other model files used for the problems described in this thesis, and files containing the experimental data used are available from the author.

```
# declaration part

Variables :=[melAq,FM,H2O,mon,NN,di,NNN,tri,N4,tet,pen,hex];
Parameters:=[fa1,E1,fam1,Em1,fa2,E2,fam2,Em2,
             FM1,FM2,FM3,FM4,FM5,FM6,FM7,FM8];
Constants :=[temp,R,tt0,tt1,mel0,mel1,begin,
             iFM1,iFM2,iFM3,iFM4,iFM5,iFM6,iFM7,iFM8];

# initial settings

Cdefault[temp] := 323;
Cdefault[R]    := 8.34;
Cdefault[tt0]  := 0.0;
Cdefault[tt1]  := 5.0;
Cdefault[mel0] := 0.12;
Cdefault[mel1] := 0.124;
Cdefault[begin]:= 1.0;
Cdefault[iFM1] := 1.0;
Cdefault[iFM2] := 0.0;
Cdefault[iFM3] := 0.0;
Cdefault[iFM4] := 0.0;
Cdefault[iFM5] := 0.0;
Cdefault[iFM6] := 0.0;
Cdefault[iFM7] := 0.0;
Cdefault[iFM8] := 0.0;

# scaling factor and reference temperature

fac :=1000.0;
Tref:=333;

# for the reparametrisation

RT1:=1.0/C[temp]-1.0/Tref;
RT2:=1.0/C[temp]-1.0/Tref;
RT3:=1.0/C[temp]-1.0/Tref;
RT4:=1.0/C[temp]-1.0/Tref;
```

```

# reparametrised reaction rates

k1 := exp(-P[fa1 ]-P[E1 ]*fac/C[R]*RT1);
km1:= exp(-P[fam1]-P[Em1]*fac/C[R]*RT2);
k2 := exp(-P[fa2 ]-P[E2 ]*fac/C[R]*RT3);
km2:= exp(-P[fam2]-P[Em2]*fac/C[R]*RT4);

# for the linear interpolation for dissolved melamine

melbeg:= C[mel0];
melend:= C[mel1];

# the corresponding algebraic equation (g[melAq]=0).

g[melAq]      := melbeg+(melend-melbeg)*(t-C[tt0])/(C[tt1]-C[tt0])-Y[melAq];

# the differential equations

f[FM]      := -k1*Y[FM]*(6.0*Y[melAq] + 4.0*Y[mon] + 2.0*Y[di] +
                    4.0*Y[NN] + 2.0*Y[NNN] + 2.0*Y[N4]) -
                k2*Y[FM]*(Y[mon] + 2.0*Y[di] + 3.0*Y[tri] +
                    Y[NNN] + 2.0*Y[tet] + Y[pen]) +
                km1*Y[H2O]*(Y[mon] + 2.0*Y[di] + 3.0*Y[tri] +
                    Y[NNN] + 2.0*Y[tet] + Y[pen]) +
                km2*Y[H2O]*(2.0*Y[NN] + 2.0*Y[NNN] + 2.0*Y[tet] +
                    4.0*Y[N4] + 4.0*Y[pen] + 6.0*Y[hex]);

f[H2O]      := -f[FM];
f[mon]      := 6.0*k1 *Y[FM]*Y[melAq]+2.0*km1*Y[H2O]*Y[di]+
                2.0*km2*Y[H2O]*Y[NN]-4.0*k1 *Y[FM]*Y[mon]-
                k2 *Y[FM]*Y[mon]-km1*Y[H2O]*Y[mon];
f[NN]       := k2*Y[FM]*Y[mon]+km1*Y[H2O]*Y[NNN]-
                4.0*k1*Y[FM]*Y[NN]-2.0*km2*Y[H2O]*Y[NN];
f[di]       := 4.0*k1*Y[FM]*Y[mon]+3.0*km1*Y[H2O]*Y[tri]+
                2.0*km2*Y[H2O]*Y[NNN]-2.0*k1*Y[FM]*Y[di]-
                2.0*k2*Y[FM]*Y[di]-2.0*km1*Y[H2O]*Y[di];
f[NNN]      := 4.0*k1*Y[FM]*Y[NN]+2.0*k2*Y[FM]*Y[di]+
                4.0*km2*Y[H2O]*Y[N4]+2.0*km1*Y[H2O]*Y[tet]-
                k2*Y[FM]*Y[NNN]-2.0*k1*Y[FM]*Y[NNN]-
                2.0*km2*Y[H2O]*Y[NNN]-km1*Y[H2O]*Y[NNN];
f[tri]      := 2.0*k1*Y[FM]*Y[di]+2.0*km2*Y[H2O]*Y[tet]-
                3.0*k2*Y[FM]*Y[tri]-3.0*km1*Y[H2O]*Y[tri];
f[N4]       := k2*Y[FM]*Y[NNN]+km1*Y[H2O]*Y[pen]-
                2.0*k1*Y[FM]*Y[N4]-4.0*km2*Y[H2O]*Y[N4];
f[tet]      := 3.0*k2*Y[FM]*Y[tri]+2.0*k1*Y[FM]*Y[NNN]+
                4.0*km2*Y[H2O]*Y[pen]-2.0*k2*Y[FM]*Y[tet]-
                2.0*km2*Y[H2O]*Y[tet]-2.0*km1*Y[H2O]*Y[tet];

```

```

f[pen] := 2.0*k2*Y[FM]*Y[tet]+2.0*k1*Y[FM]*Y[N4]-
          km1*Y[H2O]*Y[pen]-4.0*km2*Y[H2O]*Y[pen]+
          6.0*km2*Y[H2O]*Y[hex]-k2*Y[FM]*Y[pen];
f[hex] := k2*Y[FM]*Y[pen]-6.0*km2*Y[H2O]*Y[hex];

# initial conditions (different for every series)

YStart[melAq]:=0.12*C[iFM1]+0.14*C[iFM2]+0.11*C[iFM3]+
              0.17*C[iFM4]+0.25*C[iFM5]+0.3*C[iFM6]+
              0.15*C[iFM7]+0.17*C[iFM8];
YStart[FM] :=P[FM1]*C[iFM1]+P[FM2]*C[iFM2]+P[FM3]*C[iFM3]+
              P[FM4]*C[iFM4]+P[FM5]*C[iFM5]+P[FM6]*C[iFM6]+
              P[FM7]*C[iFM7]+P[FM8]*C[iFM8];
YStart[H2O] :=34.0;
YStart[mon] :=0.0;
YStart[NN] :=0.0;
YStart[di] :=0.0;
YStart[NNN] :=0.0;
YStart[tri] :=0.0;
YStart[N4] :=0.0;
YStart[tet] :=0.0;
YStart[pen] :=0.0;
YStart[hex] :=0.0;

# used to estimate the relative error during the numerical integration

YSize[melq]:=10.0;
YSize[FM] :=10.0;
YSize[H2O] :=10.0;
YSize[mon] :=10.0;
YSize[NN] :=10.0;
YSize[di] :=10.0;
YSize[NNN] :=10.0;
YSize[tri] :=10.0;
YSize[N4] :=10.0;
YSize[tet] :=10.0;
YSize[pen] :=10.0;
YSize[hex] :=10.0;

# lower bounds for the unknown parameters

ParMin[fa1] :=0.0;
ParMin[E1] :=0.0;
ParMin[fam1] :=0.0;
ParMin[Em1] :=0.0;
ParMin[fa2] :=0.0;

```

```
ParMin[E2]      :=0.0;
ParMin[fam2]    :=0.0;
ParMin[Em2]     :=0.0;
ParMin[FM1]     :=0.0;
ParMin[FM2]     :=0.0;
ParMin[FM3]     :=0.0;
ParMin[FM4]     :=0.0;
ParMin[FM5]     :=0.0;
ParMin[FM6]     :=0.0;
ParMin[FM7]     :=0.0;
ParMin[FM8]     :=0.0;

# upper bounds for the unknown parameters

ParMax[fa1]     :=5.48;
ParMax[E1]      :=196000/fac;
ParMax[fam1]    :=9.36;
ParMax[Em1]     :=136000/fac;
ParMax[fa2]     :=16.3;
ParMax[E2]      :=240000/fac;
ParMax[fam2]    :=18.98;
ParMax[Em2]     :=180000/fac;
ParMax[FM1]     :=16.82;
ParMax[FM2]     :=15.22;
ParMax[FM3]     :=11.2;
ParMax[FM4]     :=11.16;
ParMax[FM5]     :=9.6;
ParMax[FM6]     :=9.62;
ParMax[FM7]     :=9.6;
ParMax[FM8]     :=11.16;
```

Appendix 6.B

This appendix contains an example taken from a part of a data file as it is needed for the parameter estimation program, spIds ([EHS95]). The data file goes with the model file as given in Appendix 6.A. We only show the parts of the data file which are relevant to get insight into the preparation of a more complex data file and will not fill the pages with all the measurements.

```
#
# Name of the data file, start at t=0.0 and initiation of the constants
#
DATASET testDSM
START 0.0 exp1
CONSTANT temp 323.0
CONSTANT R 8.34
CONSTANT mel0 0.12
CONSTANT mel1 0.124
CONSTANT tt0 0.0
CONSTANT tt1 5.0
CONSTANT iFM1 1.0
CONSTANT begin 1.0
#
# Measurements for the various species, e.g. FM can be replaced by the
# the number 2; its position in the list variables.
#
0.0 FM 8.14
5.0 FM 8.754
5.0 mon 0.230
5.0 NN 0.223
5.0 di 0.291
5.0 NNN 0.56
5.0 tri 0.181
5.0 N4 0.031
5.0 tet 0.123
5.0 pen 0.0007
5.0 hex 0.00011
#
# Handling of the first discontinuity at t=5.0, name of the experiment
# part is prt11 and setting of the constant for this exp. part.
#
CONTINUE 5.00 prt11
CONSTANT mel0 0.124
CONSTANT mel1 0.132
CONSTANT tt0 5.0
CONSTANT tt1 15.0
CONSTANT iFM1 1.0
```

```

CONSTANT  begin  0.0
15.0  FM      4.810
15.0  mon     0.287
.      .      .
.      .      .
.      .      .
#
# end of the first experiment, start of the second experiment
#
120.0  pen     0.570
120.0  hex     0.065
STOP 120.01
START 0.0  exp2
CONSTANT  temp      323.0
CONSTANT  R          8.34
CONSTANT  mel0       0.14
CONSTANT  mel1       0.141
CONSTANT  tt0        0.0
CONSTANT  tt1        5.0
CONSTANT  iFM2       1.0
CONSTANT  begin     1.0
0.0  FM      9.26
.      .      .
.      .      .
# end of the 8th experiment and the data file.
#
120.0  pen     0.214
120.0  hex     0.012
STOP 120.01

```

Chapter 7

Software Design and Implementation

7.1 Introduction

A substantial part of the work in this PhD-project is spent on the development of a tool for parameter estimation. As a result we have built a program package called *spIds*, which is the acronym of ‘simulation and *parameter Identification in dynamical systems*’. This program enables the user to (1) simulate dynamical systems, (2) to validate models, (3) to estimate unknown parameters in such systems when additional data from experiments about the system are known and (4) to get information about the reliability of the model and the estimated parameters. In order to make the software convenient to use, we extended the number of requirements by adding that the four above points should be realised in an environment that (5) is interactive, (6) is easy to use and (7) shows the results by direct visualisation.

In this chapter we give a description of the structure of the software, the major considerations that were taken into account and the decisions we made with respect to the construction of the software. By dynamical systems we mean systems of semi-explicit differential algebraic equations (DAEs), as introduced in (1.1).

In Section 7.2 we will give an outline of the design principles of the software, whereas its structure is presented in Section 7.3. The model equations are provided by the user via the *modelfile*. This file should meet certain specifications as discussed in Section 7.4. The experimental data are made available through the *datafile*, the corresponding characteristics and specifications of this file can be found in Section 7.5.

After the model dependent parts (model and data) have been explained, we concentrate in Section 7.6 on the algebraic engine, which puts the modelfile into appropriate subroutines. The filter which takes care of handling the data is outlined in Section 7.7. Section 7.8 is devoted to the numerical engine: the part which contains numerical routines for solving the model and variational equations, optimising the criterion function and performing statistical analyses. The graphical user interface (GUI) and the database manager are the topics of the Sections 7.9 and 7.10, respectively. The last section of this chapter, Section 7.11, contains concluding remarks.

7.2 Design principles of the application spIds

The main purpose of the program is to solve a *parameter estimation problem*. I.e., the program can be used to validate mathematical models of physical (chemical, biological, biochemical etc.) processes and compute the values of unknown parameters that appear in the description of these processes (cf. Chapter 6 and e.g. [BS92, BDB86, Hem72b]). Of course, in order to determine such parameters, an unambiguous description of the model describing the process should be available. In addition, sufficient experimental data are needed, and we assume that such data are available.

We assume that the process can be modelled by a system of ordinary differential equations (*ODEs*) or a system of differential algebraic equations (*DAEs*). In fact, we assume that the process is described by an initial value problem (*IVP*) for a system of differential equations:

$$\begin{aligned}\frac{dy}{dt} &= f(t, y, \theta), \\ y(t_0, \theta) &= y_0(\theta),\end{aligned}\tag{7.1}$$

or, including the algebraic equations, by the system

$$\begin{aligned}\frac{du}{dt} &= f(t, u, v, \theta), \\ 0 &= g(t, u, v, \theta), \\ u(t_0, \theta) &= u_0(\theta).\end{aligned}\tag{7.2}$$

Here the vector $y(t, \theta) = \begin{pmatrix} u(t, \theta) \\ v(t, \theta) \end{pmatrix}$ represents the variables in the model, which describe the state of the system for $t > t_0$. In the case of the differential algebraic equations the vector $y(t, \theta)$ comprises two parts, $u(t, \theta)$ and $v(t, \theta)$. For each state variable in the first part, $u(t, \theta)$, a differential equation is available. For each remaining variable an algebraic equation is given. Of course, all *state variables* in $y(t, \theta)$ are a function of time, $t \geq t_0$, and they depend on the (unknown) parameters θ . The function $y(t, \theta)$ is called the *state vector*, as it describes the state of the physical process at time t .

To make the description of the program easier in this chapter we slightly adapt the notation of Chapter 1. In this chapter we write $f_i(t, y, \theta)$ ($1 \leq i \leq \text{nodq}$), for a differential equation and $g_j(t, y, \theta)$ ($1 \leq j \leq \text{noaq}$)¹, for an algebraic equation. With this notation we do not have to introduce the diagonal matrix A of (1.1). This has certain advantages when specifying the model in Section 7.4. The notation $u_i(t, \theta)$ for a differential variable and $v_j(t, \theta)$ for an algebraic variable such that $y(t, \theta) = (u(t, \theta), v(t, \theta))^T$ suggests that the order of the differential and algebraic variables is fixed. However, as in the model description the elements of the state vector are not numbered but identified by names,

¹Throughout this chapter we use the `typewriter` font for reserved names, that are usually denoted by a single symbol in mathematical notation.

the ordering is not relevant. Also the order of the differential and algebraic variables is not substantial.

In the description below, also the symbols for the dimensions n , m , N and K are replaced by `noq`, `nop`, `nobs` and `nosid`, respectively. The dimensions of $u(t, \theta)$ and $v(t, \theta)$ (cf. (7.2)) are denoted by `nodq` and `noaq`, respectively, such that `nodq` + `noaq` = `noq`. The system of ODEs, (7.1), can be seen as a special case of the system of DAEs, with `noaq` = 0.

To solve the differential equations, an initial vector $u(t_0, \theta)$ should be given. The program requires to provide a complete initial state $y(t_0, \theta)$. If algebraic equations are present (`noaq` > 0), this initial state should (approximately) satisfy the conditions determined by these algebraic conditions. The initial state, $y(t_0, \theta)$, i.e. the state vector at $t = t_0$, may depend on the parameter vector θ . The number of `noq` initial values (independent initial relations) determines a unique solution of the system of DAEs (ODEs).

symbol	meaning	dimension
t	time, the independent variable	1
y	the state vector, $y = (u, v)^T$	<code>noq</code> (n)
u	the vector of state variables for which a differential equation is given (a part of y)	<code>nodq</code>
v	the vector of state variables for which no differential equation is given (a part of y)	<code>noaq</code>
θ	the vector of unknown parameters	<code>nop</code> (m)
C	a vector of known constants	<code>noc</code>
f	a vector function of t , y and θ , that describes the rate of change of u with respect to t .	<code>nodq</code>
g	a vector function of t , y and θ , that describes the algebraic relations between the components of y .	<code>noaq</code>
y_0	the initial condition of the DAEs (possibly depending on θ)	<code>noq</code> (n)
R	the (possibly nonlinear) constraints on θ	<code>nosid</code> (K)

Table 7.1: Summary of the symbols in the model (between brackets in the last column the symbols as used in Chapter 1)

The initial-value problem (7.1) or (7.2) is supposed to give a relevant mathematical description of the process under consideration. The set of equations (7.2), together with its initial values and possible constraints for the parameters, we call the *model*. Generally, we assume that lower and upper bounds for the unknown parameters are known, i.e. the parameter vector satisfies:

$$\theta_{min} \leq \theta \leq \theta_{max} .$$

Often there are additional *constraints* for the unknown parameters, as introduced in (1.27). The dimension of the vector $R(\theta)$ is `nosid`.

Besides a vector of unknown parameters we introduce a vector of known *constants*. This vector is denoted by C and has dimension `noc`. These constants are used for known quantities, that are constant during some part of the experiment, but may vary over different parts of the experiment. More details about the use of these constants are given in Section 7.4.

Starting the parameter estimation program, the user gets control over this application by means of the *graphical user interface (GUI)*. This means that the user gets some kind of a *dashboard* on the computer screen, and by moving the mouse and clicking the buttons he can steer the actions of the program. The GUI will show the results and it will take care of proper file management, call the necessary numerical routines and show the solution by visualisation on the screen.

Before a numerical experiment can be performed, the user has to supply the model and the measurements. This information should be provided on two files: the *modelfile* and the *datafile*. The *modelfile* contains a description of the DAEs plus the initial values and the restrictions on the parameters, the *datafile* contains the measurements.

7.3 Structure of the software

In order to get a maintainable piece of software, the structure of *spIds* is modular. In Figure 7.1, we show a schematic view of its separate parts. The parts are discussed in the following sections.

The kernel of the system is the *numerical engine*, which performs all numerical computations: it integrates the system of DAEs, performs the optimisation and analyses the final estimate statistically. This part of the system is written in FORTRAN. In order to solve a parameter estimation problem, subroutines are required to specify the problem. Of course, these subroutines are different for each model. Hence, these subroutines are generated automatically by a separate module, the *algebraic engine*, which is written in the MAPLE V language. This module only requires the description of the problem: model equations, initial conditions and optional restrictions on the parameters. By computer algebra, the algebraic engine derives the required formulae and generates the corresponding FORTRAN subroutines. Thus, it delivers the model-dependent part of the FORTRAN source for the numerical engine. The source description of the model equations, together with the initial conditions and the parameter constraints are provided by the user and put into the *modelfile*. In order that the algebraic engine will be able to handle this information properly, it should satisfy a number of specifications that are described in Section 7.4.

Besides the model description, the numerical engine requires the data from the experiments. Such data should also be given on a separate file, called *datafile*, according to certain specifications that are described in Section 7.5. These data are checked for consistency and prepared for the numerical engine by another module called the *filter*.

This filter is partly written in NAWK (for text handling) and partly in FORTRAN. Necessary data from the model to give a proper interpretation of the data on the datafile are provided by the algebraic engine to the filter through the file *names*.

After a check on the consistency of the experimental data, the data are stored in the *database*, ready for use by the numerical engine.

A third module in the system is the *graphical user interface*, or *GUI*. The tasks of this module can be divided into two parts: (i) interaction of the user with the system to steer the computational process and (ii) visualisation of results from the numerical engine.

The communication between the different modules described above is taken care of by a fourth module, the *database manager*. By means of exchanging special messages, it regulates the flow of data between the database and the other modules. Section 7.10 gives a brief description of the database manager.

The modular structure has two additional advantages. First, separate modules can run at different machines and second, the application without the GUI is still of use when no machine with sophisticated graphics facilities is available. At this moment the GUI only runs, in combination with a UNIX operating system and X-windows, on a Silicon Graphics machine, while all the other parts can be run on almost every machine which has MAPLE and a FORTRAN77 compiler.

7.4 The modelfile

The *modelfile* contains the mathematical description of the process that will be analysed. In this section we describe how the model should be specified in the modelfile.

The modelfile is written in the MAPLE language and it will be interpreted by the MAPLE program. This means that the user has the disposal of the complete MAPLE language to express his problem in a mathematical form. However, generally only a very small part of the language is necessary to specify the differential(-algebraic) equations, the initial conditions and the few other data that are necessary to formulate the model.

First, we specify the contents of the modelfile as far as it will be understood by the algebraic engine. This is done by enumerating the building parts of the modelfile and by indicating whether the parts are obligatory or optional. Second, we give a template of the modelfile in Table 7.3. Besides the typical lines that are found in the modelfile, the user is free to use additional MAPLE language to help the mathematical formulation of the problem. As in Section 7.2 the `typewriter` font is used to indicate reserved words, Table 7.2 contains a list of these words and some of their properties.

The lines that appear in the modelfile are used in order to:

1. Define the **list of state variables** as `Variables`. This list corresponds to the names of the components of the state vector, $y(t, \theta)$. Instead of the variables y_1, \dots, y_{noq} , the user is free to choose names that are more meaningful for the problem at hand.

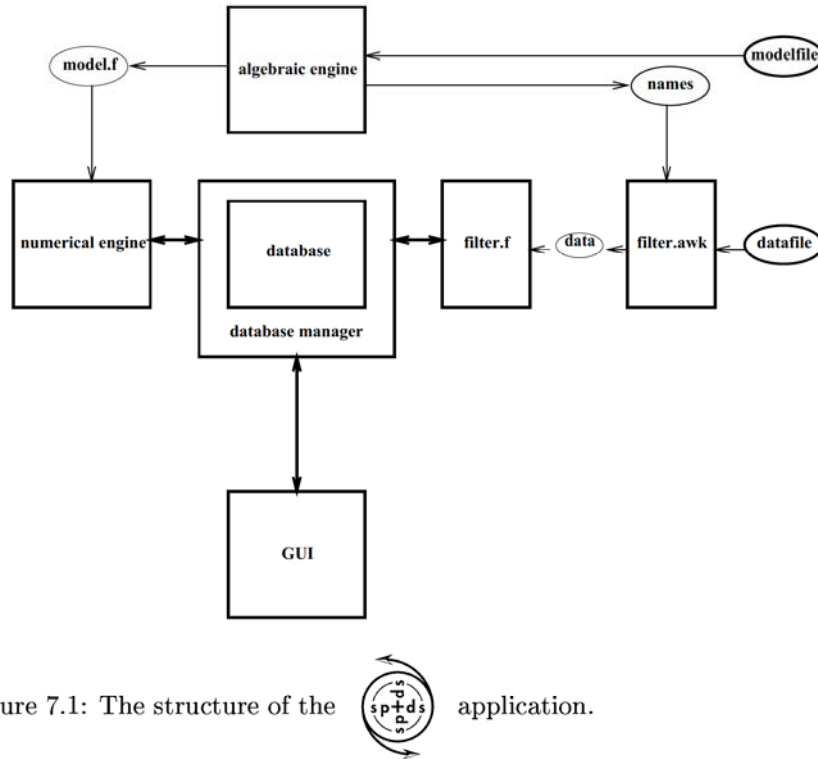



Figure 7.1: The structure of the  application.

The names of the actual variables (e.g. `vari`, $1 \leq i \leq \text{noq}$) are free for the user to choose. The number of variables, `noq`, is known to the program by the length of the list `Variables`.

2. Define the **list of unknown parameters** as `Parameters`, this list corresponds with the vector θ . The names of the parameters (e.g. `parj`, $1 \leq j \leq \text{nop}$) are free for the user to choose, but they should be different from the variable names. The number of parameters, `nop`, is known to the program by the length of the list `Parameters`.
3. Define optionally a **list of constants** as `Constants`. The length of this list will be identified as `noc`. The list contains the names for constants, introduced by the user, and gives the opportunity to identify quantities that have a fixed value for one (part of an) experiment, but that may be different (but still fixed) in another (part of the) experiment.

The names of the actual variables are free for the user to choose, but they should be different from the parameter and variable names.

If such constants are introduced in the `modelfile`, each constant should be initialised

by the user with a (default) value. For each constant (e.g. named `conk`) this is done by assigning a value to `Cdefault[conk]`. In the datafile the user will have the opportunity to overwrite these values with different values for particular (parts of) experiments. In Section 7.5 we shall see how these constants can be used.

4. Define optionally a **list of constraints** as `SideConditions` that should be satisfied by the parameter values. The length of this list will be identified by `nosid`. The list contains a name for each constraint of the form $R_i(\theta) \leq 0$, that is specified by the user (see (1.27)). Besides these additional (possible nonlinear) constraints that are specified by the user, we have constraints of the form

$$\theta_{\min} \leq \theta \leq \theta_{\max},$$

to indicate a feasible box region of the parameters.

All names introduced in the above lists should be unique names, appearing only once in all four lists.

5. Define the **right-hand sides**, $f(t, y, \theta)$, of the differential equations in (7.1) or (7.2), by assigning an algebraic expression (depending on the available `Y[vari]`, `P[parj]`, and `C[conk]`) to the array elements `f[varl]`, for $1 \leq l \leq \text{nodq}$.
6. Define the **algebraic equations**, $g(t, y, \theta) = 0$, of the DAEs by assigning an algebraic expression (depending on the available `Y[vari]`, `P[parj]`, and `C[conk]`) to the array elements `g[varl]`, for $\text{nodq} + 1 \leq l \leq \text{noq}$.
7. Define the **initial states**, $y_0(\theta)$, in (7.1) or (7.2) by assigning an algebraic expression (depending on `P[parj]`, and `C[conk]`) to the array element `YStart[vari]`, for $1 \leq i \leq \text{noq}$. It is necessary to assign expressions to all possible `YStart[vari]`, for $1 \leq i \leq \text{noq}$.

If the user forgets one of the above, required assignments, he will receive an error message. More assignments are optional. In the case that an optional specification is omitted, the program will use a default setting as given in Table 7.2.

8. Determine the `nop`-dimensional rectangle in parameter space, where the unknown parameter vector resides. In the modelfile lower- and upper-bounds for the parameter values can be given. Therefore arrays `ParMin` and `ParMax` are introduced, for which

$$\text{ParMin}[\text{parj}] \leq \theta_j \leq \text{ParMax}[\text{parj}]; \quad j = 1, \dots, \text{nop}.$$

If no values for `ParMin` and `ParMax` are specified, the default values `ParMin[parj] = 0` and `ParMax[parj] = 1` are assumed, for $j = 1, \dots, \text{nop}$.

9. Define the additional **constraints**, that were introduced in **SideConditions**. These additional (possibly nonlinear) constraints in the parameter space are specified by assigning the expressions $R(\theta)$, as in equation (1.27), to the array of expressions $R[sid1]$, with $l = 1, \dots, nosid$, where the index $sid1$ is the name in the list of the l -th parameter constraint. We call these additional constraints *side conditions*. Such expressions only depend on the unknown parameters $P[parj]$ and the known constants $C[conk]$.
10. Indicate the order of magnitude for the components in the state vector, so that

$$\max_{t, \theta} |y_i(t, \theta)| \leq YSize[vari]; \quad i = 1, \dots, noq.$$

These $YSize$ -values are used for scaling purposes only and play a minor role in the computations. If no $YSize$ is specified, its elements are assumed to be equal to 1.0.

A modelfile template

In Table 7.3 we give a template of the MAPLE-text on the modelfile. The choice of most names used in the MAPLE text are at the user's discretion, except for the reserved words as listed in Table 7.2. In this table, $var1$, $vari$, $varnoq$, $par1$, $parj$, $parnop$, $con1$, $conk$, $connoc$, $sid1$, $sidl$ and $sidnosid$, are names that can be selected by the user; $RHSexpressionj$, $ALGexpressioni$ are algebraic expressions depending on the independent variable t , the dependent variables $Y[vari]$, the parameters $P[parj]$, and the constants $C[conk]$ (with $i=1, \dots, noq$, $j=1, \dots, nop$, $k=1, \dots, noc$). The $RHSexpressionj$ corresponds with $f_j(t, y, \theta)$, ($j = 1, \dots, nodq$), and describes the right hand side of the j -th differential equation. The i -th algebraic equation, $g_i(t, y, \theta)$, ($i = 1, \dots, noaq$), is represented by $ALGexpressioni$. For all $i = 1, \dots, noq$, $INITexpressioni$ corresponds with the initial condition, $y_i(t_0, \theta)$, of the i -th component of the differential-algebraic equations, and it may depend on $P[parj]$ and $C[conk]$. The assignments to $Cdefault$, $Ysize$, $ParMin$ and $ParMax$ are expressions for numerical values (floating or fixed point numbers).

Besides the standard constraints, $\theta_{min} \leq \theta \leq \theta_{max}$, additional model constraints with respect to the unknown parameters can be added at the end of the modelfile. These side conditions, which are allowed to be nonlinear, are supplied in the form $R[sid1] := SIDEexpressionl$ (with $l=1, \dots, nosid$). Here, $SIDEexpressionl$ is an algebraic expression, depending on the unknown parameters $P[parj]$ and the known constants $C[conk]$, representing the expression $R_l(\theta)$, the l -th component in equation (1.27). The number of side conditions ($nosid$) corresponds with the dimension of $R(\theta)$. We assume, as we do for $INITexpressioni$, that $SIDEexpressionl$ is (MAPLE-) differentiable with respect to θ . For $RHSexpressionj$ and $ALGexpressioni$ we assume (MAPLE-) differentiability with respect to θ and to y .

A complete example of a modelfile is found in Appendix 6.A.

Reserved name	Assignment	Default value	Type in MAPLE
Variables	yes		list ^a
Parameters	yes		list ^a
Constants	optional		list ^a
SideConditions	optional		list ^a
t	no		name
Y	no		table ^b
P	no		table ^b
C	no		table ^b
f	yes		table ^c
g	for DAEs		table ^c
R	optional	R[sid1]=-1.0, (1 ≤ l ≤ nosid)	table ^d
YStart	yes		table ^d
YSize	optional	YSize[vari]=1.0, (1 ≤ i ≤ noq)	table ^e
ParMin	optional	ParMin[parj]=0.0, (1 ≤ j ≤ nop)	table ^e
ParMax	optional	ParMax[parj]=1.0, (1 ≤ j ≤ nop)	table ^e
Cdefault	optional	Cdefault[conk]=0.0, (1 ≤ k ≤ noc)	table ^e

^alist of names

^btable of variables

^ctable of expressions, depending on Y, P and C

^dtable of expressions, depending on P and C

^etable of floating point numbers

Table 7.2: Summary of reserved names and default values in the modelfile

7.5 The datafile

The datafile contains the measured values (*observations*) obtained from the process studied. From equation (1.2) we see that the measured value, \tilde{y}_i , is related to the point in time t_i and the component c_i of the state vector, $1 \leq c_i \leq \text{noq}$. In the datafile all information about a single measured value should be given on one single line. So the *data part* of the simplest datafile consists of `nobs` lines, with on each line three numbers: t_i , c_i and \tilde{y}_i . The numbers t_i and \tilde{y}_i are floating point numbers, c_i is an integer that corresponds with the c_i -th variable in the list of variables. This integer can also be replaced by the

```

Variables:=[var1,vari,varnoq];
Parameters:=[par1,parj,parnop];
Constants:=[con1,conk,connoc];
SideConditions:=[sid1,sidl,nosid];
Cdefault[con1]:= constant1;
Cdefault[conk]:= constantk;
Cdefault[connoc]:= constantnoc;
f[var1]:= RHSExpression1;
g[vari]:= ALGexpressioni;
f[parj]:= RHSExpressionj;
g[noq]:= ALGexpressionnoq;
YStart[var1]:= INITexpression1;
YStart[vari]:= INITexpressioni;
YStart[varnoq]:= INITexpressionnoq;
YSize[var1]:= ysize1;
YSize[vari]:= ysizei;
YSize[varnoq]:= ysizeoq;
ParMin[par1]:= parmin1;
ParMin[parj]:= parminj;
ParMin[parnop]:= parminnop;
ParMax[par1]:= parmax1;
ParMax[parj]:= parmaxj;
ParMax[parnop]:= parmaxnop;
R[sid1]:= SIDEexpression1;
R[sidl]:= SIDEexpressionl;
R[sidnosid]:= SIDEexpressionnosid;

```

Table 7.3: The template of a modelfile.

corresponding symbolic name that appears in the list **Variables**. The lines corresponding with a single experiment should appear in the order of increasing (more precisely: nondecreasing) t_i .

A single experiment

In the simplest possible datafile, the *data part* is preceded by two lines: (1) a line containing some identification of this data set: an arbitrary string of at most 24 characters, and (2) a line containing only the word **START** and the value t_0 . This obligatory line denotes that the initial value problem starts at $t = t_0$. The data part is closed by a single line, containing the word **STOP** and the value for t_{end} , the time at which the initial value problem ends.

If the user wants to provide a weight w_i for the weighted sum of squares (1.4), he

can do this by adding the real number w_i as the 4-th number on the line for the i -th observation. If no weight is specified, it has the same effect as $w_i = 1.0$.

In case the user wants to skip a measurement, he can inactivate the measurement by putting a 0 as the 5-th number at the end of the corresponding line. We give this number the name *active*. The default setting is 1, which means that the measurement is *active*, i.e. is taken into account during the computation.

Multiple experiments

Another important option is to take several experiments into account for the same model and the same parameters, but possibly with different values of the model constants as given in *constants*. In this manual we use the word *experiment* for a sequence of observations (measurements) ordered in time. In case of a parameter estimation problem with a series of experiments the user should provide a series of *data parts*, each of which is preceded by a line containing the value t_0 (to denote that a new initial value problem is considered, starting at $t = t_0$). In order to specify what values for the constants are used, the restart line can be immediately preceded by a number of lines which contain the word *CONSTANT*, the constant's name, and the constant's value.

In this way, a datafile can contain measurements from many different experiments corresponding to the same modelfile. If some constants change from one experiment to the other, the corresponding measurements have to be separated by a constant block.

It is also possible to change the model constants at distinct values t_{cont} within the range of integration, $t_0 < t_{\text{cont}} < t_{\text{end}}$. At such times, t_{cont} , a discontinuity in the process of the experiment occurs and the change of constants is specified in the datafile: e.g. an amount of a certain reactant is added during the experiment or the temperature changes. So, each experiment may consist of different, distinct periods, where the constants have fixed values. Such periods during experiments are called *experiment parts*.

At the beginning of every experiment the constants are set equal to their default values from the modelfile and adaptation will be made after every appearance of a constant-line in the datafile.

Datafile syntax

Summarising we find the following syntax for the information on the datafile.

```

DATA_FILE:          identification_line ; EXPERIMENT_BLOCK
EXPERIMENT_BLOCK:  EXPERIMENT [ ; EXPERIMENT_BLOCK ]
EXPERIMENT:        START_PART [ ; CONTINUATION ] ; stop_line
CONTINUATION:      CONTINUATION_PART [ ; CONTINUATION ]
START_PART:        start_line [ ; CONST_PART ] ; DATA_PART
CONTINUATION_PART: continue_line [ ; CONST_PART ] ; DATA_PART
CONST_PART:        constants_line [ ; CONST_PART ]
DATA_PART:         data_line [ ; DATA_PART ]
identification_line: DATASET , data-set-name
start_line:        START ,  $t_0$  [ , experiment-name ]
continue_line:     CONTINUE ,  $t_{\text{cont}}$ 
stop_line:         STOP ,  $t_{\text{end}}$ 
constants_line:    CONSTANT , conk ,  $b_j$ 
data_line:          $t_i$  ,  $c_i$  ,  $\tilde{y}_i$  [ ,  $w_i$  [ , 0 | 1 ] ]
comment_line:      #, a sequence of characters ending with carriage return

```

In this syntax description, ‘;’ means ‘followed on the next line by’,
‘,’ means ‘followed on the same line by’, ‘[]’ means ‘optional’,
and ‘|’ means ‘or’.

t_i , \tilde{y}_i , w_i and b_j are floating point numbers;
 c_i is a natural number;
 c_i can be replaced by `vari` from the list `Variables` ;
`conk` is an element from `constants` in the model.

data-set name and experiment name are sequences of at most 24 characters. The binary flag (0|1) on the data line denotes that the observation is active. All data lines in an experiment block, following a "START ; t_0 ; name"-line, should be ordered in time such that $i < j \Rightarrow t_i \leq t_j$. Also the possible t_{cont} should satisfy this ordering. Any such sequence beginning with a t_0 is called an *experimental sequence* and can be identified by an experiment-name.

A partial example of a datafile, containing the essential parts, is given in Appendix 6.B.

7.6 Algebraic Engine

A modelfile which satisfies the specification of Section 7.4 can be used as input for the algebraic engine. A schematic overview of the interaction of the algebraic engine with its environment and its position in the overall application is given in Figure 7.1. The algebraic engine generates the model-dependent part of the numerical engine. These

model-dependent parts are written to the file *model.f*. This file contains FORTRAN sub-routines for the evaluation of: (i) the differential algebraic equations (1.1), (ii) their derivatives with respect to the state variables and the parameters (cf. (1.6)), (iii) the initial conditions, (iv) their derivatives with respect to the parameters, (v) the discontinuities within the experiments, (vi) the restrictions on the parameters and (vii) their derivatives with respect to the parameters ((1.27) and (1.28), respectively). The choice here for FORTRAN is motivated by the fact that its use is widely spread and can be easily integrated with robust, public domain numerical software routines. Both arguments make it available for a broader group of users. A disadvantage of FORTRAN is the memory allocation, which should be handled via hard upperbounds which have to be adjusted manually if the problem sizes exceed an a priori chosen maximum.

The other part of the output of the algebraic engine is the file *names*. This file contains the number of variables, parameters, constants and restrictions and their corresponding, user supplied names. The numbers of each of them are of interest for the array bound checks of the memory allocation. The names will be checked with the names of the constants and variables which are present in the datafile, and transferred to the database afterwards.

7.7 Filter

The purpose of the filter is twofold. First, it checks whether the datafile matches the modelfile. Second, it puts the measurements in the database. For the first purpose we start with checking the format of the data file; whether it meets the specifications of Section 7.5. Subsequently, we use the file *names*, which was created by the algebraic engine, to check the consistency between the modelfile and the datafile. All the constants and state variables in the datafile should be present in the modelfile. This part of the filter is written in NAWK.

For the second purpose, another part, *filter.f*, is written and checks the array bounds for the number of measurements, experiment-parts and experiments. If one of these bounds is exceeded this filter gives an error message and the bounds for the memory allocation should be adapted. If the filter handles the datafile without error messages, then the information from the *datafile* is properly located in the central database.

7.8 Numerical engine

The numerical engine takes care of (i) the computation of an approximate solution of the model equations and the corresponding variational equations, 1.1 and 1.5, respectively, (ii) minimising the criterion function (e.g. (1.4), (2.4), (3.11) or (3.55)) and (iii) performing statistical analyses and investigate the nonlinearity, cf. Section 1.6 and Chapter 4.

The numerical engine can be divided into two parts; (i) a part which depends on the parameter estimation problem and (ii) a problem-independent part. The problem-

dependent part, *model.f*, comes from the algebraic engine as described in Section 7.6 and is linked together with the model-independent part of the numerical engine. The result is a binary program which performs the numerical work. Due to its tasks, the numerical engine is the most CPU time consuming part of the whole application.

The problem-independent part also has a modular structure in itself for the same reasons as the whole application has a modular structure. The separate parts of the numerical engine will be highlighted in the subsequent paragraphs. The measurements and constants from the datafile, which are necessary for the numerical engine, are obtained via the central database manager.

A special BDF solver, which exploits the stiffness structure of the variational equations as described in Section 1.3, forms one of the modules of the numerical engine. This solver uses the model dependent part, *model.f*, for evaluation of the right-hand sides of (7.1) or (7.2) and their derivatives. During the calculation, not only the model responses which correspond to the measurements are calculated, but also model responses for visualisation purposes are calculated. Every time the initial value problem is solved for a vector of parameters these results are sent to the GUI via the DB manager for direct visualisation.

After the BDF solver calculated the discrepancies (1.9) and the Jacobian (1.11) for a given value of θ , the Levenberg-Marquardt minimisation routine (Section 1.5) will do one step in this iterative process in order to find an improved parameter vector which gives a smaller value for the least squares criterion, subsequently the model and variational equations will be solved with this improved parameter vector. Iterative procedures for total least squares, maximum likelihood or L_1 -estimates will be dealt with in a similar way. The minimisation part also uses the model dependent part, *model.f*, for evaluation of the parameter constraints (1.27) and their derivatives (1.28).

Upon convergence of the Levenberg-Marquardt algorithm, linear statistical analyses in the vicinity of the calculated optimal estimated parameter vector, $\hat{\theta}$, as described in Section 1.6, are performed. Intersections of the ellipsoidal confidence with the parameter axes, as shown in Figure 1.1, can be studied via the GUI in combination with the corresponding SVD decomposition.

Derivation of nonlinearity measures as described in Chapter 4 is another part of the numerical engine. One of the options is to compare ellipsoidal regions, as they can be derived from linear theory, with the corresponding results from Monte Carlo simulations. This and the other nonlinearity measures are tools to enable the user to investigate the nonlinearity of the problem under consideration.

7.9 Graphical user interface (GUI)

The graphical user interface (GUI) is designed in order (i) to make the whole application interactive in an easy way and (ii) to have the option to view the results immediately during the computation. The first item covers starting/stopping a computation, adjust input parameters of the numerical routines (accuracy, maximum number of iterations),

change upper and lower bounds for the parameters to be estimated, change the model file or the datafile. The second item concerns the visualisation of graphs of the best fit at that moment of the computation, follow a track in the parameter space during the minimisation and graphical representation of results of the statistical analysis.

The part of the GUI which is relevant for the user is a ‘dashboard’ with buttons, scroll-down menus, viewers and sliders to enable the user to steer easily through the options and be able to use the software with hardly any instructions. It is also designed to change the problem formulation slightly, adapt the numerical accuracies, view numerical results, without typing long command lines, but pushing these buttons with the mouse and opening submenu’s instead. This idea makes it much faster and easier to perform the many tasks due to immediate interfering with the computational process, without typing and without consulting the manual, because the user interface is partially self explaining.

The results from the model investigations are visualised on the screen immediately and can be stopped by the user at any moment in order to change the initial parameters or its bounds, adjust numerical accuracy, adapt the data or even switch to a more sophisticated model. C.T.H. Everaars built the GUI, more details about the concept and the realisation can be found in [EHS95].

7.10 Database manager

This central part of *spIds* takes care of the communication between the different parts of the software package. By means of events it regulates the flow of data between the database and the other modules or satellites.

A copy of all data that can be communicated between the different modules is put in the database. Therefore all information is stored double, which not only minimises communication when multiple processors are used, but also forms a backup if one of the parts stops or communicated data get lost due to external errors. Besides, it stores actual numerical results from the numerical engine and it delivers them to the GUI for visualisation, if required.

The content of the database is grouped on behalf of their characteristics and communication frequencies. For instance, tasks –a special kind of communicated data– for the numerical engine from the GUI are put in one group. The content of each group can be changed by at least one module and this changed content is of interest for at least one other part. The database manager takes care of a proper administration of these events.

The database manager is designed by R. van Liere, the general concept of this manager can be found in [Lie92, WL96], for more technical details see [LW96].

7.11 Concluding remarks

In this chapter we started by giving a list of requirements for a parameter estimation tool, the relevance for such a tool is motivated by problems encountered in experimental sciences as described in Chapter 6. From these requirements we derived a top-bottom design for a modular setup of the software. The choices and decisions we made at the various stages and levels of the design have been motivated throughout the chapter. With respect to the input of the program –the model and the data– we gave detailed and precise specifications to obtain an unambiguous formulation. Much attention is paid to error detection in the program input.

Problem-depended software is generated automatically, by using computer algebra, and merged with numerical routines needed to solve parameter estimation problems in dynamical systems. The tool is completed with a graphical user interface which makes it interactive, easy to use and enables the user to see the numerical results immediately. The data communication between the different modules is taken care of by a database manager. This choice for the communication keeps the overall application modular and enables the user to run different modules on different machines.

Samenvatting

Parameters in de beschrijving van tijdsafhankelijke fysische processen zijn grootheden met een praktische relevantie, omdat ze het model een vrijheid geven die gebruikt kan worden om de beschrijving met waargenomen feiten overeen te laten stemmen. Dikwijls zijn deze parameters onbekend en niet rechtstreeks te bepalen. Een manier om ze te schatten wordt in dit proefschrift behandeld. De dynamische systemen die we hier beschouwen worden gemodelleerd met behulp van differentiaal-algebraïsche vergelijkingen: de modelvergelijkingen. Om de parameters te kunnen schatten, is een aantal metingen nodig dat betrekking heeft op de toestandsvariabelen van de onderliggende modelvergelijkingen. De oplossing van deze vergelijkingen is afhankelijk van de parameters, die zodanig worden gekozen dat de metingen en de uitkomsten van het model zoveel mogelijk met elkaar in overeenstemming zijn.

Verschillende criteria kunnen worden gebruikt om de mate van deze overeenstemming te quantificeren. De keuze van zo'n criterium hangt af van de kennis en de aannames met betrekking tot de meetfouten. Aanvankelijk zullen we uitgaan van normaal verdeelde fouten met een gegeven covariantiematrix, hetgeen aanleiding geeft tot kleinste-kwadraten (OLS) schatters. Als de onafhankelijke variable (de tijd) ook onderhevig is aan meetfouten, dan gebruiken we de totale kleinste-kwadraten (TLS) methode. De OLS en TLS methoden worden behandeld in het eerste gedeelte van dit proefschrift. Hier ligt het accent op de numerieke aanpak van parameterschattingsproblemen in dynamische systemen. Daarnaast worden de betrouwbaarheidsgebieden van de geschatte parameters bepaald en niet-lineaire restricties op de parameters beschouwd.

Normaal verdeelde meetfouten met een onbekende covariantie matrix geven aanleiding tot de zogenaamde meest aannemelijke (ML) schatters, die algemener zijn dan kleinste-kwadraten schatters. Wanneer een methode voor kleinste kwadraten reeds beschikbaar is, dan valt de berekening van de ML schatters relatief eenvoudig te implementeren. Hetzelfde geldt als de meetfout een Laplace verdeling heeft en de L^1 -norm van de fout geminimaliseerd wordt. Een bijkomend voordeel van de laatstgenoemde schatters is dat ze robuuster zijn, d.w.z. minder gevoelig voor uitschieters, dan kleinste-kwadraten schatters.

Na deze verhandeling over schatters worden de belangrijkste verschillen tussen lineaire en niet-lineaire regressie besproken, gevolgd door een beschrijving van verschillende methoden om de mate van niet-lineariteit uit te kunnen drukken. De niet-lineariteit kan gesplitst worden in een intrinsieke en een parameter afhankelijke niet-lineariteit. De eerste hangt af van de modelvergelijkingen en de gekozen experimentele opzet. De

tweede hangt af van de parametrisatie van het model en kan verminderd worden indien een geschikte herparametrisatie toegepast kan worden. In geval van een geringe mate van niet-lineariteit kan men voor de berekening van de betrouwbaarheidsgebieden van de parameter schattingen volstaan met een lineaire benadering. Is daarentegen de niet-lineariteit substantieel dan moeten rekenintensieve methoden aangewend worden om de parameterruimte in kaart te brengen.

Aansluitend wordt de aandacht verlegd naar het ontwerpen van optimale experimentopzetten. Hierbij gaat het om het plannen van vervollexperimenten, met het oog op modeldiscriminatie, het verkleinen van de betrouwbaarheidsgebieden van de geschatte parameters of het reduceren van de niet-lineariteit.

De beschreven theorie wordt toegepast op een scala aan praktijkproblemen uit de industrie, (bio-)chemie, econometrie en op een testprobleem uit de literatuur. Voor elke toepassing is enige kennis op het gebied van het onderliggende fysische proces onontbeerlijk om tot een succesvol eindresultaat te komen.

Gedurende het gehele onderzoek hebben de theoretische en rekenkundige aspecten enerzijds en de problemen uit de praktijk anderzijds een sterke wisselwerking op elkaar uitgeoefend. De verbinding tussen beide heeft geresulteerd in de totstandkoming van een softwarepakket voor het oplossen van parameterschattingsproblemen. De beschreven theorie en de bijbehorende implementatie is getoetst aan de hand van tal van praktijkproblemen. Door de opzet van de software is het mogelijk om verschillende onderdelen parallel op verschillende workstations te laten werken. Voor de communicatie tussen de software en de gebruiker is een speciale grafische interface ontwikkeld.

Bibliography

- [And68] J.F. Andrews. A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrates. *Biotechnol. Bioeng.*, 10:707–723, 1968.
- [Bar74] Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, New York and London, 1974.
- [BBS87] P.T. Boggs, R.H. Byrd, and R.B. Schnabel. A stable and efficient algorithm for nonlinear orthogonal distance regression. *SIAM J. Sci. Stat. Comput.*, 8(6), 1987.
- [BCC⁺92] C. Bischof, A. Carle, G. Corliss, A. Griewank, and P. Hovland. ADIFOR - Generating derivative codes from FORTRAN programs. *Scientific Programming*, 1(1):11–29, 1992. Also retrievable through:
<http://www.mcs.anl.gov/adifor/AdiforDocs.html>.
- [BD87] P.J. Brockwell and R.A. Davis. *Times Series: Theory and Methods*. Springer-Verlag, New York, 1987.
- [BDB86] L.T. Biegler, J.J. Damiano, and G.E. Blau. Nonlinear parameter estimation: a case study comparison. *AIChE Journal*, 32(1):29–45, 1986.
- [BF90] P.J. Brown and W.A. Fuller, editors. *Statistical Analysis of Measurement Error Models and Applications*. Contemporary Mathematics, Volume 112. American Mathematical Society. Providence, Rhode Island, 1990.
- [BG97] P.B. de Bruin and J.G. de Gooijer. A comparison of ARMA and SETAR forecasts. Preprint, submitted to *Sankhyā*, 1997.
- [BHA64] G. E. Becker, P.A. Hui, and P. Albersheim. Synthesis of extracellular polysaccharides by suspension of *Acer pseudoplatanus* cells. *Plant Physiol.*, 39:913–920, 1964.
- [BKMA74] D.P. Burke, M. Kaufman, M. McNeil, and P. Albersheim. The structure of plant cell walls. VI. a surgery of the walls of suspension-cultured monocots. *Plant Physiol.*, 54:109–115, 1974.
- [BN89] C.M. Bailey and H. Nicholson. A new structured model for plant cell cultures. *Biotechnol. Bioeng.*, 34:1331–1336, 1989.
- [Boc85] H.G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität, Bonn, 1985.
- [Box71] M.J. Box. Bias in nonlinear estimation. *J.R. Statistical Soc., Ser. B*, 33(2):171–201, 1971.
- [BS92] E. Baake and J.P. Schlöder. Modelling the fast fluorescence rise of photosynthesis. *Bulletin of Math. Biology*, 54(6):999–1021, 1992.

- [BTH⁺95] E.G. Bovill, R.P. Tracy, T.E. Hayes, R.J. Jenny, F.H. Bhushan, and K.G. Mann. Evidence that meizothrombin is an intermediate product in the clotting of whole blood. *Arterioscler. Thromb. Vasc. Biol.*, 15(6):754–758, 1995.
- [BW80] D.M. Bates and D.G. Watts. Relative curvature measures of nonlinearity. *J. R. Statist. Soc. B*, 42(1):1–25, 1980.
- [BW88] D.M. Bates and D.G. Watts. *Nonlinear Regression Analysis and its Applications*. John Wiley & Sons, Inc., 1988.
- [CGG⁺91] B.W. Char, K.O. Geddes, G.H. Gonnet, B.L. Leong, M.B. Monagan, and S.M. Watt. *Maple V Library Reference manual*. Springer Verlag, 1991.
- [CK86] M.F. Chaplin and J.F. Kennedy. *Carbohydrate analysis, a practical approach*. IRL Press Limited, Oxford, 2nd edition, 1986.
- [DS81] N.R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, Inc., 2nd edition, 1981.
- [DS83] J.E. Dennis, Jr. and R.B. Schnabel. *Numerical Mathematics for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, N.J., 1983.
- [Efr79] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 11:1–27, 1979.
- [EHS95] C.T.H. Everaars, P.W. Hemker, and W. Stortelder. *Manual of splds, a software package for parameter identification in dynamic systems*. Technical Report NM-R9521, CWI, Amsterdam, 1995.
- [EJ68] J. Edelman and T.G. Jeford. The mechanism of fructosan metabolism in higher plants as exemplified in *Halianthus tuberosus*. *Plant Physiol.*, 93:148–161, 1968.
- [EW95] L. Edsberg and P.-Å. Wedin. Numerical tools for parameter estimation in ODE-systems. *Optimization Methods and Software*, 6:193–217, 1995.
- [Fed72] V.V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.
- [Fra89] G.C. Frazier. A simple leaky cell growth model for plant cell aggregates. *Biotechnol. Bioeng.*, 33:313–320, 1989.
- [Ful87] W.A. Fuller. *Measurement Error Models*. John Wiley & Sons, Inc., 1987.
- [Gea71] C.W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, Inc. Englewood Cliff, NJ, 1971.
- [GHW66] M. Gordon, A. Halliwell, and T. Wilson. Kinetics of the addition state in the melamine-formaldehyde reaction. *Journal of Applied Polymere Science*, 10:1153–1170, 1966.
- [Gle90] L.J. Gleser. Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In P.J. Brown and W.A. Fuller, editors, *Statistical Analysis of Measurement Error Models and Applications*, Contemporary Mathematics, Volume 112, pages 99–114. American Mathematical Society. Providence, Rhode Island, 1990.
- [GR86] D. Grindlay and T. Reynolds. The *aloe vera* phenomenon: A review of the properties and modern uses of the leaf parenchyma gel. *J. Endopharmacol.*, 16:117–151, 1986.

- [GV83] G.H. Golub and C.F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1983.
- [Hem72a] P.W. Hemker. Numerical methods for differential equations in system simulation and in parameter estimation. In H.C. Hemker and B. Hess, editors, *Analysis and Simulation of Biochemical Systems*, pages 59–80. North Holland Publ. Comp., 1972.
- [Hem72b] P.W. Hemker. *Parameter estimation in nonlinear differential equations*. Technical Report MR 134, Mathematical Centre, Amsterdam, 1972.
- [Hem93] H.C. Hemker. Thrombin generation, an essential step in haemostasis and thrombosis. In A.L. Bloom and D. Thomas, editors, *Haemostasis and thrombosis, 3E*, pages 477–491, 1993.
- [HK93] P.W. Hemker and J. Kok. *A project on parameter identification in reaction kinetics*. Technical Report NM-R9301, CWI, Amsterdam, 1993.
- [HNW93] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer-Verlag, Berlin Heidelberg, second revised edition, 1993.
- [HPD87] A.D. Hale, C.J. Pollock, and S.J. Dalton. Polysaccharide production in liquid cell suspension cultures of *phleum L.* *Plant Cell Rep.*, 6:435–438, 1987.
- [Hub81] P. Huber. *Robust Statistics*. John Wiley, New York, 1981.
- [HW94] D. Hermey and G.A. Watson. Some robust methods for nonlinear parameter estimation. In D. Bainov and V. Covachev, editors, *2nd Int. Coll. on Numerical Analysis*, pages 93–102. VSP, 1994.
- [HW96] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin Heidelberg New York, second revised edition, 1996.
- [LBS92] Th.W. Lohmann, H.G. Bock, and J.P. Schlöder. Numerical methods for parameter estimation and optimal experiment design in chemical reaction systems. *Industrial & Engineering Chemistry Research*, 31:54–57, 1992.
- [Lie92] R. van Liere. Computational steering: a case study. *CWI Quarterly*, 5(3):207–217, 1992.
- [Loh93] Th.W. Lohmann. *Ein numerisches Verfahren zur Berechnung optimaler Versuchspläne für beschränkte Parameteridentifizierungsprobleme*. PhD thesis, Universität Augsburg, 1993.
- [LSV96] W.M. Lioen, J.J.B. de Swart, and W.A. van der Veen. *Test set for IVP solvers*. Technical Report NM-R9615, CWI, Amsterdam, 1996. Available via WWW at URL <http://www.cwi.nl/cwi/projects/IVPtestset.shtml>.
- [LW96] R. van Liere and J.J. van Wijk. CSE: A Modular Environment for Computational Steering. In M. Gobel, J. David, P. Slavik, and J.J. van Wijk, editors, *Proceedings of the 7th Eurographics Workshop on Visualization in Scientific Computing, Prague*, pages 256–266. Springer-Verlag, 1996.
- [MA95] J.C. Merchuk and J.A. Asenjo. Fundamentals of bioreactor design. In J.C. Merchuk and J.A. Asenjo, editors, *Bioreactor System Design*, pages 139–207. Marcel Dekker, New York, 1995.

- [Mar63] D.W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11:431–441, 1963.
- [MFG92] L.B. Marshall, N.L. Figler, and S.L. Gonias. Identification of alpha 2-macroglobulin conformational intermediates by electron microscopy and image analysis. Comparison of alpha 2-macroglobulin-thrombin and alpha 2-macroglobulin reacted with cis-dichlorodiammineplatinum(II) and trypsin. *Journ. Biol. Chem.*, 267(9):6347–6352, 1992.
- [MGB74] A.M. Mood, F.A. Graybill, and D.C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, Inc., 1974.
- [MS62] T. Murashige and F. Skoog. A revised medium for rapid growth & bioassays with tobacco tissue cultures. *Plant Physiol.*, 15:473–479, 1962.
- [MT90] D.D. Monkovic and P.B. Tracey. Activation of human factor V by factor Xa and thrombin. *Biochemistry*, 29(5):1118–1128, 1990.
- [Neu90] M. Neushul. Antiviral carbohydrates from marine red algae. *Hydrobiol.*, 205:99–104, 1990.
- [Pin95] J.D. Pintér. LGO: An implementation of a lipschitzian global optimization procedure. User's guide. Technical Report NM-R9522, CWI, Amsterdam, 1995.
- [Pol86] C.J. Pollock. Fructans and metabolism of sucrose in vascular plants. *New Phytol.*, 104:1–24, 1986.
- [PSS97] J.D. Pintér, W. Stortelder, and J.J.B. de Swart. *Computation of the elliptic Fekete points*. Technical Report MAS-R9705, CWI, Amsterdam, 1997. Submitted to *Journal of Global Optimization*.
- [Rao73] C.R. Rao. *Linear Statistical Interference and Its Applications*. Wiley, New York, 2nd edition, 1973.
- [Rat83] D.A. Ratkowsky. *Nonlinear Regression Modeling*. Marcel Dekker, Inc., New York, 1983.
- [Ray88] D. Ray. Comparison of forecasts: an emperical investigation. *Sankhyā*, 50B:258–277, 1988.
- [Sch59] H. Scheffé. *The Analysis of Variance*. John Wiley & Sons, Inc., 1959.
- [Sch85] J.P. Schlöder. *Numerische Methoden zur Behandlung hochdimensionaler Aufgaben der Parameteridentifizierung*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität, Bonn, 1985.
- [SD83] M.L. Shuler and M.M. Domach. Mathematical modeling of the growth of individual cells. In H.W. Blanch, E.T. Papoutsakis, and G. Stephanopoulos, editors, *Foundations of Biochemical Engineering*, Washington, ACS Symposium Series Number 207, pages 93–99, 1983.
- [Seb77] G.A.F. Seber. *Linear Regression Analysis*. Wiley, New York, 1977.
- [Sil80] S.D. Silvey. *Optimal Design*. Chapman and Hall, 1980.
- [ST85] H. Schwetlick and V. Tiller. Numerical methods for estimating parameters in nonlinear models with errors in the variables. *Technometrics*, 27(1):17–24, 1985.
- [SW88] G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. John Wiley & Sons, Inc., 1988.

- [UIFN74] Y. Ueda, H. Ishiyama, M. Fuki, and A. Nishi. Invertase cultured *D. carote* cells. *Phytochem.*, 13:383–387, 1974.
- [VPR96] S. Van Huffel, H. Park, and J.B. Rosen. Formulation and solution of structured total least norm problems for parameter estimation. *IEEE Transactions on Signal Processing*, 44(10):2464–2474, 1996.
- [Wat94] D.G. Watts. Estimating parameters in nonlinear rate equations. *The Canadian Journal of Chemical Engineering*, 72:701–710, 1994.
- [WB73] R.L. Whistler and J.N. BeMiller. *Industrial gums, Polysaccharides and their Derivatives*. Academic Press, New York, 1973.
- [Wei89] A. Weimken. Fructan synthesis in excised barley leaves. *Plant Physiol.*, 101:459–468, 1989.
- [Wid72] J.M. Widholm. The use of fluorescein diacetate and phenosafranin for determining viability of cultured plant cells. *Stain Tech.*, 47(4):345–351, 1972.
- [Wik97] G. Wikström. *Algorithms and software for the computation of parameters occurring in ODE-models*. UMINF 97.03. Department of Computing Science, Umeå University, S-901 87 Umeå, Sweden, 1997. Licentiate Thesis.
- [WL96] J.J. van Wijk and R. van Liere. An environment for computational steering. In G.M. Nielson, H. Müller, and H. Hagen, editors, *Scientific Visualization: Overviews, Methodologies, and Techniques*. 1996.
- [WP97] E. Walter and L. Pronzato. *Identification of parametric models from experimental data*. Springer-Verlag, Paris, 1997.
- [YS77] T.S. Yamaoka and S. Satto. *Bot. Mag. Tokyo*, 90:153–163, 1977.

Index

- L_1 -optimisation, 44
- χ^2 -distribution, 10, 26
- ϵ -rank, 64

- activation energy, 83
- amidolytic activity, 93
- aramide yarn, 122
- ARMA, 128
- Arrhenius' law, 83

- backward differentiation formulae, 7
- Barnes' problem, 15
- BDF, 7
- bias measure of Box, 53
- biomass, 107

- Cell suspension, 104
- chromogenic substrate, 94
- clotting factor, 92
- component, 5
- computer algebra, 7, 27
- confidence region, 10, 26
- constants, 152
- constrained minimisation, 12, 26
- constraints, 152
- covariance matrix
 - measurement errors (OLS), 9
 - measurement errors (TLS), 25
 - parameters, 10
- currency notes, 128

- DAE, 150
- DAEs, 4
- data part, 157–159
- datafile, 149, 157
- dependent confidence interval, 11
- deviation, 31

- differential algebraic equations, 4
- diffusion, 122
- discrepancy, 5
- double exponential distribution, 43

- econometrics, 128
- ellipsoidal confidence region, 11
- enzymatic reaction, 13
- errors in variables method, 21
- EVM, 21
- experiment, 159
 - part, 159
 - sequence, 160

- F-ratio test, 15, 18, 99
- fast equilibrium, 117
- Fisher information matrix, 50
- Fisher's F-distribution, 10, 26
- formaldehyde, 80
- fresh biomass, 109

- gas constant, 83
- Gauss-Newton, 8, 24
- Gaussian distribution, 30
- generalised least squares, 31
- GLS, 31
- gradient-based minimisation, 5
- graphical user interface, 162
- graphical user interface (GUI), 152
- grid refinement, 124
- growth associated, 107
- GUI, 162

- Henry constant, 118
- Hessian
 - of $S(\theta)$, 36
 - weighted discrepancies, 9, 53

- heteroscedasticity, 136
- Huber M-estimator, 44
- hydrolysis, 106
- improved E-design, 72
- independent confidence interval, 10
- individual confidence regions, 26
- inhibitor, 92
- intrinsic curvature, 59
- IVP, 150
- Jacobian
 - model equations, 6
 - variational equations, 6
 - weighted discrepancies, 8
- Lagrange multipliers, 12, 27
- Laplace distribution, 43
- Levenberg-Marquardt, 8, 25
- lifted line, 57
- likelihood function, 30
- LLF, 30
- Lotka-Volterra, 15
- lysis, 106
- MAPLE language, 153
- MAPLE V, 7
- marginal confidence region, 12
- MC-method, 52
- measurement, 5, 22
- melamine, 80
- method of lines, 124
- methylation, 79
- Michaelis-Menten relation, 96
- model, 151
- modelfile, 149, 153
- moment matrix, 32, 42
- Monod kinetics, 107
- Monte Carlo, 52
- Newton's method, 9
- nonlinear regression, 29
- nonviable cells, 106
- normal curvature, 58
- normal distribution, 30
- normal equations, 8, 24
- observations, 157
- ODE, 150
- OED, 65
- OLS, 3
- optimal experiment design, 65
- Ordinary Least Squares, 3
- orthogonal least squares, 21
- parameter constraints, 12, 26
- parameter-effect curvature, 58
- parameters, 4
- partial pressure, 118
- PDE, 123
- pdf, 10
- perturbation, 52
- plant tissue, 104
- polysaccharide, 105
 - extracellular, 109
 - intracellular, 108
- pre-exponential factor, 83
- predator-prey, 14
- principal components, 64
- probability density function, 10, 30
 - Laplace distribution, 44
 - TLS, 41
- reference temperature, 86
- relative normal curvature, 58
- relative parameter-effect curvature, 59
- reparametrisation, 85
- reserved words, 156
- residuals, 30
- resin, 79
- sample, 32
- satellites, 163
- saw-tooth, 128
- sensitivity equations, 5
- SETAR, 128

- side conditions, 156
- singular value decomposition, 8
- site, 54
- solution locus, 55
- spIds, 149
- standard deviation, 31
- standard radius, 58
- state vector, 4
- steady state, 123
- steepest descent, 8
- substrate, 108
- SVD, 8
- Symphytum officinale* L., 104

- titer, 123
- TLS, 21
- total least squares, 21
- trapezoidal rule, 124

- unbiased estimator
 - of σ^2 (OLS case), 10
 - of σ^2 (TLS case), 26

- variational equations, 5
- viability, 106
- viable cells, 106

- water penetration, 122
- weights
 - delayed, 35
 - frozen, 35
 - OLS, 5
 - TLS, 23

ISBN: 90-74795-91-9
Front label: Tobias Baanders
Printing and binding: Ponsen en Looijen B.V., Wageningen
Number of copies: 350