ISSN 1360-1725

THE UNIVERSITY
of MANCHESTER

UMIST

Models with Delays for Cell Population Dynamics: Identification, Selection and Analysis.

Part I: Computational Modelling with Functional Differential Equations: Identification, Selection, and Sensitivity.

Christopher T.H. Baker, Gennadii A. Bocharov, Christopher A.H. Paul & Fathalla A. Rihan

Numerical Analysis Report No. 425
Printed February 27, 2003

Manchester Centre for Computational Mathematics
Numerical Analysis Reports

DEPARTMENTS OF MATHEMATICS

Reports available from:
Department of Mathematics
University of Manchester
Manchester M13 9PL
England

And over the World-Wide Web from URLs
http://www.ma.man.ac.uk/nareports
ftp://ftp.ma.man.ac.uk/pub/narep

Contents

Models with Delays for Cell Population Dynamics: Identification, Selection and Analysis.

C.T.H. Baker[*] G.A. Bocharov[†] C.A.H.Paul[‡] & F.A. Rihan[§]

Report425.tex

Abstract

In this work, the authors aim at a unified approach to the identification of mathematical models that can be applied to studies of cell proliferation. The presentation is made in three parts. The first relates to a discussion of background modelling considerations (the purpose of a mathematical model is, in part, to promote scientific understanding and the development of the theory, to permit prediction of observable phenomena, and to influence experimental design). The second part relates to models based upon a system of differential equations with memory. The third part relates specifically to the employment of a hierarchy or family of parametrized neutral delay differential equations in modelling cell dynamics through a mechanism employing a suitable 'best fit' criterion.

For a given model equation, the term 'best fit' is defined by a suitable choice of objective function (which represents quantitative consistency); the quality of fit contributes – along with a measure of parsimony, for example – to an indicator of (in some sense) acceptability. Since different models, or the same model with different parameters, can be suggested by a given set of observations, we discuss ways of identifying suitable models and discriminating systematically between the various candidates. In particular, we discuss (under assumptions that are precise) the interplay between model evaluation criteria or "model-selection procedures", the choice of objective function, nonlinearity effects, and sensitivity analysis.

The models discussed here, which simulate experimental observations, employ suitable functional differential equations, such as delay differential equations or neutral delay differential equations, incorporating time-lags or memory effects. In an earlier paper (*Modelling and analysis of time lags in some basic patterns of cell proliferation*, [10]), to which this paper is a sequel, Baker *et al.* observed that some basic patterns of cell dynamics can be simulated accurately using equations of this type. We propound a family or hierarchy of models, incorporating scientifically meaningful parameters whose values are estimated using observed data, by optimizing a measure of best fit to the given data.

All the mathematical processes involve advanced computational procedures (in particular, numerical methods for the solution of functional differential equations) as well as – or in combination with – the more traditional analytical and statistical tools. Using numerical techniques developed by the authors, we analyze some examples of models of cell growth.

Keywords: Cell growth, model selection, neutral delay differential equation, parameter estimation, first & second order sensitivity analysis, nonlinear bias, objective functions, methodologies.

[*] cthbaker@ma.man.ac.uk Communicating author: Research Professor, Department of Mathematics, The University of Manchester, Manchester M13 9PL, England; Visiting Professor, Chester College.

[†] bocharov@inm.ras.ru Honorary Research Fellow, The University of Manchester; Leverhulme Visiting Professor, Chester College; Permanent affiliation, Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow; work performed, in part, whilst on leave at Imperial College School of Medicine, London.

[‡] chris@maths.man.ac.uk, Research Fellow, Department of Mathematics, The University of Manchester

[§] frihan@ma.man.ac.uk Honorary Research Fellow, Department of Mathematics, The University of Manchester. Permanent address: Department of Mathematics, Helwan University, Cairo, Egypt.

Models with Delays for Cell Population Dynamics: Identification, Selection and Analysis.

Part I

Computational Modelling with Functional Differential Equations: Identification, Selection, and Sensitivity.

C.T.H. Baker* & G.A. Bocharov†, and C.A.H.Paul‡ & F.A. Rihan§

February 27, 2003

Abstract

This is the first in a series of reports in which we discuss aspects of the identification of, discrimination between, and sensitivity of, mathematical models, and we illustrate with some examples.

Mathematical models based upon certain types of differential equations, functional differential equations, or systems of such equations, are often employed to represent the dynamics of natural, in particular biological, phenomena. We present some of the principles underlying the choice of a methodology for the computational development of quantitatively consistent models, using scientifically meaningful parameters, based on observational data.

Exploiting the link between a (weighted) least-squares fit, information theory, and maximum likelihood, we can select an appropriate objective function which is minimized over a set of parameters. To compare the best fit models from amongst a hierarchy of possible models we can employ indicators that can be regarded as taking parsimony into account. A sensitivity analysis provides feedback on the covariances of the parameters in relation to the best fit and on whether we need to include all the parameters in the model. We attempt to formulate a methodology for assessing the adequacy of mathematical models, founded on this approach.

Much of our thinking is influenced by discussions that are dispersed amongst the existing literature. However, we believe that the value of our material is more than the sum of that of its separate parts. Additionally, in a number of places we extend or supplement or reformulate existing results. The models that we consider generate infinite-dimensional dynamical systems.

Key words. Computation, data sampling, modelling, objective function, parsimony criteria, sensitivity, identifiability, well-posedness.

I-1 Introduction

One of the significant challenges in bio-mathematics (and other areas of science) is to formulate meaningful mathematical models[1]. We regard a model as meaningful if it is both descriptive and predictive — if it is consistent with previous observations and can predict future behaviour (e.g. the outcome of further observations or experiments) under changed conditions. In this sense, a good model will reduce the need to perform certain experiments; however, the production of a good model depends upon the availability of sufficient data of sufficient quality.

[1] This is the first in a number of reports by the authors that address various aspects of mathematical modelling.

Our main interest has been in modelling the growth dynamics of *in vitro* and *in vivo* biological systems (such as cell proliferation). Whereas our interest is focussed in this direction, the mathematical and computational strategies we employ have features to be found in other areas, and our remarks have widespread applicability.

We address ways of discriminating systematically between different models, though the use of quantitative indicators that permit a ranking of possible alternatives. In particular, we discuss topics that (while discussed in other areas) have been relatively under-represented in the context of the modelling of the life sciences (immune response, diseases, cell-proliferation). We consider identifiability, model evaluation criteria or "model-selection procedures", the choice of objective function, parsimony, nonlinearity effects, and the sensitivity of the estimates of the parameters to uncertainty in the observed data as well as inter-relations between these aspects.

The methodology that we recommend relies on weighted least squares fitting with appropriate weights. (These weights, the parameter identification, and the model selection procedure, are associated with a maximum likelihood approach.) In order to obtain appropriate weights, we require estimates of the standard deviation of the observations. These estimates are refined by a process based upon a minimization of maximum likelihood. A product of the calculation is the maximum likelihood estimate of the variation in the data. If we wish to estimate the confidence intervals for our optimum parameters, this can be achieved using the sensitivity coefficients which indicate the sensitivity of the solution to the parameters. (In Parts II and III [11, 12] we consider various other sensitivity issues in greater depth.) Our recommended methodology involves advanced computational techniques, as well as more traditional analytical and statistical approaches, and we illustrate the discussion with some examples of cell growth, using numerical techniques developed by the authors.

We wish to emphasize that robust numerical techniques for the approximate solution of differential and functional differential equations used as models (see below) produce densely defined approximations to the solutions that retain and are consistent with the accuracy of the underlying solver. We regard soundly based and 'robust' numerical techniques as essential for quantitatively consistent modelling.

The name often given to the type of model identification that we shall address is *data assimilation*. Ordinary differential equations, partial differential equations, and integral equations have all been given rôles in mathematical models of biological, chemical, and physical phenomena (for example). For background reading, we note that models based upon ordinary differential equations and upon partial differential equations are discussed in [4, 16]; and models based incorporating hereditary effects (delay, neutral, integral or integro-differential equations) are discussed in [33, 44, 40]. Where ordinary differential equations or Volterra integro-differential equations are employed as models, it is a small step to extend the models to admit delay differential equations (DDEs) and neutral delay differential equations (NDDEs). The preceding types of equation are all examples of differential equation (DE) or functional differential equation (FDE). Various types of model have their supporters, often irrespective of the context in which they are proposed, but a methodology (based on "*model averaging*") for the use of differing models, in order to provide inferences, is described by Burnham & Anderson ([24, p.326] and [25, p.448]).

Mathematical discussions found in the literature are often based upon a qualitative study of the models (which are often kept simple — perhaps artificially simple — in order to permit a mathematical analysis). This provides a different emphasis from the one that we adopt, in which we seek, using real data, to estimate parameters in models that are both qualitatively and quantitatively consistent with observations.

I-1.1 Objectives of the discussion

We will consider methodologies for the identification of, discrimination between, and sensitivity of, mathematical models. The models we discuss are deterministic with a continuous state variable. Model identification includes, but is not limited to, the estimation of parameters in various mathematical models. We are motivated by the use of (systems of) FDEs for use in modelling cell

proliferation, but we are led to consider, in a broader context, various issues of model identification and we expect our observations to be of wider relevance.

As Kolmanovskii and Myskhis write, in their book [38],

> "The basic view on scientific modeling is that a model is any 'simplified description of a system (etc.) that assists calculations and predictions' (Oxford English Dictionary). The main reason to model something is to provide for an efficient organization of information and experiences in order to enhance understanding and enable (wise) decision making."

Mathematical models can be based on various types of mathematical systems. Many scientists have considered the difficulties associated with mathematical modelling. As Hopkins & Leipold [37] observed:

> "Mathematical models can be roughly divided into empirical and mechanism based models. Empirical models typically comprise an arbitrary mathematical function and suitable parameter values that adequately describe the process being modelled." ..... "Mechanism-based models, on the other hand, attempt to describe a system in terms of identifiable physical processes."

From our perspective, both the underlying science and an understanding of the transient and long-term dynamics associated with different types of model equations should inform the choice of possible models. In general, one seeks model parsimony[2] that is consistent with scientific understanding and experimental data, and there should be an interaction, which flows in every direction, between the mathematical modelling and the science and the experimental work. It is our view that a multi-disciplinary approach is essential.

In [10], various levels of complexity were provided by formulating a hierarchy of cell-growth models based upon different classes of equations incorporating *time-lags*, each requiring the specification of scientifically meaningful parameters. Appleton, in a presentation on which [5] is based, provided a list of fundamental questions related to parametrized models of cell proliferation but which are equally relevant to more general modelling problems. The list included:

- How should the best fit be assessed?
- How should the model equations be solved?
- Which model is better: that having the fewer parameters or that based more closely on the biology of the system?
- Is it possible to distinguish between the important and the unimportant assumptions in a model?

The preceding remarks relate to general modelling strategies and philosophies that underpin our thinking, though we shall not address them specifically. The models we considered in [10], and consider in Parts II and III [11, 12], have a rational basis (motivated by theoretical understanding of the relevant science and of the mathematics) and are parameterized by scientifically meaningful parameters. Estimates of these parameters are computed, using experimental observations, by optimizing a measure of 'best' fit to the given data where 'best' is defined by (numerical) minimization of an objective function. There are varying motivations for choosing particular types of objective function, and in our discussion we highlight a unifying theme that leads us to our recommended methodology. In our view, employing the 'ranking' indicators discussed here – including measures of sensitivity – provides criteria for the selection of an appropriate model from a family of models, and a way to understand the dynamics of the underlying biological system.

This report was initiated as a response to the need to discriminate between different models. Much of our thinking is influenced by discussions that are dispersed amongst the existing literature. However, we believe that the value of our material is more than the sum of that of its separate parts. Additionally, in a number of places we extend or supplement or reformulate existing results (notably, existing results for classical non-linear regression and results based upon scalar – univariate – modelling).

[2] Parsimony may be described as the sparing use of resources. A dictionary definition of the law of parsimony is that no more should be assumed than is necessary to account for the facts. The introduction of a time-lag is in some sense an example of parsimony: a time-lag (for example, in gestation that proceeds to term) frequently substitutes for complex dynamical processes that would otherwise require separate modelling.

I-2 Fundamental Issues Concerning Modelling

The general discussion here will be illustrated by reference to models based upon ODEs, DDEs, and NDDEs. The computational approach permits the choice of realistic and, if necessary, quite complex, equations. The ODEs, DDEs and NDDEs that we later (*cf* Parts II & III [11, 12]) consider as potential models have solutions that we denote $\mathbf{y}(t) = \mathbf{y}(t;\mathbf{p}) \in \mathbb{R}^M$, with parameter $\mathbf{p} \in \mathbb{R}^L$; the models have one of the following forms:
ODEs:

$$\begin{array}{rcl} \mathbf{y}'(t;\mathbf{p}) & = & \mathbf{f}(t,\mathbf{y}(t;\mathbf{p})), \quad \text{for } t \in [t_0,T]; \\ \mathbf{y}(t_0;\mathbf{p}) & = & \mathbf{y}_0(\mathbf{p}); \end{array} \tag{I-2.1a}$$

DDEs:

$$\begin{array}{rcl} \mathbf{y}'(t;\mathbf{p}) & = & \mathbf{f}(t,\mathbf{y}(t;\mathbf{p}),\mathbf{y}(t-\tau;\mathbf{p})), \quad \text{for } t \in [t_0,T]; \\ \mathbf{y}(t;\mathbf{p}) & = & \boldsymbol{\psi}(t;\mathbf{p}), \quad \text{for } t \in [t_0-\tau,t_0]; \end{array} \tag{I-2.1b}$$

(where $\tau \geq 0$) and NDDEs:

$$\begin{array}{rcl} \mathbf{y}'(t;\mathbf{p}) & = & \mathbf{f}(t,\mathbf{y}(t;\mathbf{p}),\mathbf{y}(t-\tau;\mathbf{p}),\mathbf{y}'(t-\tau;\mathbf{p});\mathbf{p}), \quad \text{for } t \in [t_0,T]; \\ \mathbf{y}(t;\mathbf{p}) & = & \boldsymbol{\psi}(t;\mathbf{p}), \quad \mathbf{y}'(t;\mathbf{p}) = \boldsymbol{\psi}_1(t;\mathbf{p}), \quad \text{for } t \in [t_0-\tau,t_0]; \end{array} \tag{I-2.1c}$$

(where $\tau \geq 0$, and if $\tau > 0$ it represents a time-lag). Here, the *form* of $\mathbf{f}$ is known (and $\mathbf{f}$ is defined precisely if the parameters are specified); $\boldsymbol{\psi}$ and $\boldsymbol{\psi}_1$ are initial functions and for a given choice of parameter $\mathbf{p}$ the solution values $\mathbf{y}(t_j;\mathbf{p})$ with components $y^i(t_j;\mathbf{p})$ will be expected to simulate the observed data $\{\mathbf{y}_j\}$ with components $\{y_j^i\}$ $(i = 1,2,\ldots,M,\ j = 1,2,\ldots,N)$.[3] We can identify parametrized initial conditions as well as parametrized equations.

Remark I-2.1 Our remarks are capable of extension to models based on partial differential equations and integro-differential equations. There is, indeed, a link between models based upon DDEs and certain hyperbolic PDEs [20].

It may be asked where the mathematical models originate. There is a vast body of expertise devoted to the principles of applied mathematical modelling. The present authors would counsel against an approach based on what is described in one paper as "bits of mathematical machinery that behave in accordance with what is known about a system without constituting any sort of explanation of the behaviour". The objective is to write down "appropriate" mathematical equations and any boundary conditions, or constraints. It would require a monograph to review the principles of applied mathematical modelling. In the context of biomathematics, we may cite [3, 4, 9, 16, 44, 51, 61] as examples of good practice. For example, the book [3] addresses issues of compartmental modelling, [61] contains valuable biological information and reflections on modelling philosophy within theoretical regulatory biology.

We need to consider the type of model used (the equations combined with the boundary conditions form the mathematical model), the principles behind model identification, and computational methods for estimating parameters via a choice of objective function. We reserve to later a discussion of whether the problem that we face is well-posed.

I-2.1 Data

We presume that we have available observations that we take as input data. It is possible that data to be found in an existing paper comes from a single experiment, but it is more likely that the data arises from several experiments conducted in slightly different environments. In this case, one expects the data to represent the mean (interpreted, consistently, either as the arithmetic mean or as the geometric mean) of the measured quantities. Since the basis of the mathematical modelling usually includes assumptions about the distribution (*e.g.*, normal or log-normal, etc.)

[3] We remark that with little amendment we can consider the case where different components $\{y_j^i\}_{j=1}^{N_i}$ could be associated with i-dependent arguments t_j^i.

of the errors in the observations, it is important to ensure that these assumptions are realistic. The (*a posteriori*) sensitivity analysis may indicate that some observations (such as those relating to transient phenomena) have more impact than others on the identification of a suitable model, and this information may affect the design of the experiment. We would expect that one outcome of the mathematical modelling would be an input into the design of the experiments, in particular the data sampling strategy. Given that mathematical modelling improves over time, publication of the raw data and a precise statement of the conditions under which it was obtained will be a welcome improvement on past practice.

In those cases where observations are presented as multiple sets of data (not as the mean of differing sets of observations), there are elementary statistical formulae for estimating the variances $\{\sigma_i^j\}^2$ about the mean values y_i^j at time t_j. In this circumstance, one strategy is choose the parameter $\mathbf{p}$ in the model defining $\mathbf{y}(t;\mathbf{p})$ in order to minimize

$$\Phi(\mathbf{p}) := \sum_{i,j} \frac{1}{\{\sigma_i^j\}^2} |y^i(t_j;\mathbf{p}) - y_j^i|^2.$$

As elsewhere, this objective function represents the square of a norm of the distance of the vectors $\mathbf{y}(t_j;\mathbf{p})$ from the vectors of observations $\mathbf{y}_j$ associated with a certain set of observation times $\{t_j\}$.

There are echos of the above procedure in what we propose as a methodology, but we are proceeding on the basis that we are provided with the data in the form of the mean values, not the individual observations. In consequence, we shall (*inter alia*) need a soundly-based technique for estimating the variances. In general, the parameters being estimated will comprise $p_1, p_2, \ldots, p_L$ and the set of values σ_i^j; frequently we assume that σ_i^j is independent of i and j so that a single value σ characterises all the variances.)

I-2.2 Model identification

Model identification may be regarded as having two components: (*i*) selection from a range of forms of parametrized mathematical model, and (*ii*) computation of the values of the parameters that yield a (numerically computed) best fit. These two components are inter-related, and an attempt to reconcile them can be sought using modern theories of informational complexity (see, for example, [46]). The perspectives that one may adopt are affected by the nature of the data.

Apart from a few allusions to statistics, much of the discussion in Section I-3.1 below, has the appearance of a deterministic treatment. In reality this is not the case: There are (broadly speaking) three approaches to model selection, each of which impinges on the choice of an objective function. The approaches are the *information theoretic approach*, the *maximized likelihood approach*, and the *weighted least-squares data-fitting approach*. There is a theme running through all these approaches that (given appropriate assumptions and for a specified set of mathematical models) prompts a choice of objective function and a measure that may be taken as a guide to model selection (the choice of a model from a collection of candidate models) in order to construct, given the purpose of the model, a ranked hierarchy.

If the data is subject to observational errors of a precise statistical nature (independent, having Gaussian distribution, and zero mean) — concepts made precise in [41, Eqns(4),(5)] — we are led to statistical arguments; however, it will transpire that maximum likelihood estimates have a rôle in guiding our choice of objective function. By this route, an apparently deterministic technique of minimizing an objective function is seen to be consistent with the information theoretic approach and the maximum likelihood approach [24, Chapter2]. (Burnham and Anderson [25, p. 98] consider the theory and practical use of the information-theoretic approach to model building and data analysis to be simpler than that of the Bayesian approaches.)

I-2.3 Well-posedness and identifiability

It is often said that parameter estimation is ill-posed. Let us recall the concepts of well-posedness and ill-posedness, introduced at the beginning of the last century, by J. Hadamard. A *well-posed problem* has one and only one solution depending continuously on the data; an ill-posed

problem is one that is not well-posed[4] It appears that Hadamard held the opinion that physical situations always lead to *well-posed* problems; today this view is challenged, but it may be remarked that a mathematical formulation of a problem can sometimes be ill-posed precisely because the mathematical model does not include all scientifically relevant features.

We observed that parameter estimation is often regarded as ill-posed but, as a mathematical statement, this sentiment requires elaboration, not least in choosing between the scenarios itemized in (i) to (v) below. Thus we may distinguish, in model identification, the cases where we have

(i) a prescribed form of parametrized model where the aim is to determine numerical values of scientifically meaningful parameters (the assumption being that the form of the model is accepted as appropriate);

(ii) a hierarchy of models of differing complexity, for each of which we seek appropriate parameters. The aim here is to select a particular, parametrized, model (having a sound scientific basis and a degree of parsimony) that agrees qualitatively and quantitatively with observations;

and data that is in the form of

(iii) mathematically defined entities (e.g., observations that are presented as an error-free densely-defined function – an idealized case);

(iv) a discrete set of observational values (e.g., observations that are presented as an error-free function defined on a *discrete* set of points that may be either predetermined or modifiable – an idealized case);

(v) an aggregate of observations, on a discrete set of points, that may be subject to errors (errors that are, say, normally distributed - an idealized but frequently assumed case);

and so on. Thus, there may be no set of parameters that provides a model solution that agrees everywhere with a densely-defined function representing observed data, but there may be a unique set of parameters that provides a "most acceptable fit" to a collection of discrete data.

Towards one end of the spectrum is the definition adopted by Sjoerd Verduyn Lunel [59, Abstract]:

> "*Parameter identifiability* is concerned with the question whether the parameters of a specific model can be identified from knowledge about certain solutions of the model, assuming perfect data."

(That the data is discretely defined is subsequently apparent.) From the perspective adopted in our paper, identifiability is more concerned with whether there is an identifiable model from amongst a hierarchy of models that optimises some measurement of best fit to noisy data (a measurement that may include an index reflecting complexity or parsimony). Whilst the analysis of identifiability within one scenario may provide insight in the context of a different scenario, one should be clear about which problem is under consideration at any given time.

I-2.4 Well-posedness

We shall make some observations about well-posedness before examining (see §I-2.6) the nature of our specific problem in the context of identifiability.

The fundamental question governing any mathematical problem is whether it is well-posed in the sense of Hadamard: does a solution exist, is it unique, and if it exists does it depend continuously upon the data. The notion of continuity invoked in the term "continuous dependence" indicates that ill-posedness can only be discussed within a precise mathematical framework which includes

[4] The classical example of an ill-posed problem is a Fredholm integral equation of the first kind (slightly more generally, the solution of an equation $Kf = g$ for f, g in a Banach space X where K is a linear compact operator from X into X).

definitions of the data and solution spaces. There is a distinction between ill-posedness and ill-conditioning. In an ill-conditioned problem, the unique solution may depend continuously on the data but small changes in the data may give rise to relatively large changes in the solution; one can quantify the degree of ill-conditioning by introducing a *condition number*. Sometimes, the term "ill-posed" can also be quantified[5], and one can introduce the notions of "severely ill-posed problems" or "mildly ill-posed problems". The discretization of an ill-posed problem often results in an ill-conditioned problem in which the degree of ill-conditioning depends on the mode of discretization.

Remark I-2.2 Various regularization techniques for ill-posed problems can be located in the literature [58]; these have the effect of replacing an ill-posed problem by a nearby problem that is well-posed. A number of these regularization techniques can be associated with a least-squares approach, in the context of the present discussion, one may wish to amend the choice of objective function ($\Phi_{OLS}(\{t_\ell\}_0^N;\mathbf{p})$, $\Phi_{\mathcal{W}LS}(\{t_\ell\}_0^N;\mathbf{p})$, etc.) by minimizing a new objective function formed by the addition of a regularization term that depends upon a regularization parameter. See, for example, Baker & Parmuzin [13]. For regularization and data assimilation in other contexts see [43].

I-2.5 Theoretical identifiability

The subject of *a priori* identifiability of models is generally addressed in the context of "system-experiment models", under ideal conditions of an *error-free model structure* and *noise-free observations* [7, 19]. Thus the assumption is that the form of the model (which is characterized by a vector of parameters $\mathbf{p}$) has been correctly identified, and the observational data $\{\mathbf{y}_j\}$ is correct.

Let us examine the following examples.

Example I-2.1 *A simple example is provided by considering $y'(t) = p_0 y(t)$ for $t \geq 0$, with $y(0) = p_1$, which has solution $y(t) = p_1 \exp(p_0(t))$. If $y(1)$ is observed, then $q_1 = p_1 \exp(p_0)$ has the uniquely determined value $q_1 = y(1)$ but p_0 and p_1 are not separately identifiable.*

Example I-2.2 *Now consider a simple model governed by the NDDE*

$$\begin{array}{ll} y'(t) = \rho_0 y(t) + \rho_1 y(t-1) + \rho_2 y'(t-1), & t \geq 0 \\ y(t) = \psi(t), \quad y'(t) = \psi'(t), & t \leq 0. \end{array} \tag{I-2.2}$$

Suppose that we have values $y_i \equiv y(t_i)$ for $i = 0, 1, \ldots N$ where $t_i \in [0,1]$ and $\psi(t) \equiv \alpha$. Then, on $[0,1]$,

$$y'(t) = \rho_0 y(t) + \alpha\rho_1.$$

On $[0,1]$ we have $y(t) = \dfrac{\alpha\rho_1}{\rho_0}\{\exp(\rho_0 t)\{1+\dfrac{\rho_0}{\rho_1}\}-1\}$. That is, the model is not uniquely identifiable for arbitrary data defined only *on $[0,1]$; we can at best (if $N \geq 1$ then 'best' is indeed possible) determine the two values ρ_0 and the* product *$\alpha\rho_1$ rather than the separate parameters α, ρ_0, ρ_1. If additional data is provided on the interval $[1,2]$, this particular problem is resolved [59].*

In the latter example, we see the introduction of an "observational parameter" $\mathbf{q}$ with components $q_0 = \rho_0$ and $q_1 = \alpha\rho_1$ which defines an infinite number of parameters $\mathbf{q} = [\rho_0, \rho_1, \alpha]^T$. The general question is whether one can proceed, and if so in a *unique* way, to determine observational parameters $\mathbf{q}$ from $\{\mathbf{y}_j\}$ and thence to obtain $\mathbf{p}$ from $\mathbf{q}$. The possibilities include: (*i*) the non-existence of suitable parameters; (*ii*) the existence of a unique parameter; (*iii*) the existence of finitely many (non-unique) parameters; (*iv*) the existence of infinitely many parameters.

Cobelli and his co-authors [7, 8, 28] categorize types of identifiability with these possibilities in mind. *A priori* identifiability is regarded, ideally, as a prerequisite for parameter estimation. Assessing *a priori* global identifiability, for nonlinear or even linear models, is however difficult, as it requires solving a system of nonlinear equations which increases both in nonlinearity and

[5] For first-kind Fredholm integral equations on the space of square-integrable functions, the rate of decay to zero of the singular values can be taken to provide an indication of the degree of ill-posedness.

number of terms and unknowns with increasing model complexity; see [7, 8]. In our models, we endeavour to employ the model parameters themselves as the observational parameters.

We give a further example from DDEs; this shows, *inter alia* that the design of the experiment (at the least, the selection of the points $\{t_i\}$) is a factor in determining identifiability. The discussion following shows that there may not be a unique set of parameters whatever the size of the sample set.

Example I-2.3 *Consider a simple model associated with the DDE:*

$$\begin{array}{rcll} y'(t) & = & \rho_0 y(t) + \rho_1 y(t-1), & t \geq 0 \\ y(t) & = & \psi(t) & t \leq 0 \end{array} \tag{I-2.3}$$

Then with $\psi(t) = \exp(\gamma t)$ *the parameters* $\rho_0 = 1$, $\rho_1 = 1$ *and* $\gamma = 0$ *(*$\psi(t) = 1$*) we obtain the unique solution* $y(t) = 2\exp(t) - 1$ *on* $[0, 1]$. *On the other hand, with the parameters* $\rho_0 = 0$, $\rho_1 = 2\mathrm{e}$ *and* $\gamma = 1$ *we obtain the same unique solution. Thus any exact data specified for an arbitrary number of arguments* $\{t_j\}_{j=1}^N$ *in the interval* $[0, 1]$ *corresponds to two possible vectors of parameters,* $[\rho_0, \rho_1, \gamma]^T$. *That is, the model is not uniquely identifiable for arbitrary data defined on* $[0, 1]$.

I-2.6 Practical identifiability

The preceding discussion is based on a different perspective than that implicit in our practical approach. Our techniques for parameter estimation involve (*inter alia*) minimizing a function $\Phi_\star(\mathbf{p})$ of the parameters $\mathbf{p}$ which is constructed from the observed data. This objective function is non-negative, and therefore has at least one minimum on any compact domain of parameter values. (If the objective function vanishes at $\widehat{\mathbf{p}}$, we have an exact fit of the solution to the data when $\mathbf{p} = \widehat{\mathbf{p}}$.) However, the minimum need not be uniquely identified.

In general, for a given data set $\{t_j, \mathbf{y}_j\}_{j=1}^N$ and an arbitrary function $\mathbf{f}$ in (I-2.1), there is no reason to suppose that there exists a *unique* minimizer $\widehat{\mathbf{p}}$ of $\Phi(\mathbf{p})$. Indeed, it is easy to find examples for non-unique best fit models; one requires only to find solutions for two different parameters that agree at the points $t_1, t_2, \ldots, t_N$. In FIG. I-2.1 we give an example of such a scenario; plotting the graphs of solutions corresponding to small initial functions for the equation

$$y'(t) = y(t)\Big[a - y(t-1)\Big], \quad t \geq 0; \text{ with } y(t) = \psi(t),\ t \in [-1, 0] \tag{I-2.4}$$

where $\mathbf{p} = [a]$, and $1 \leq a \leq 1.6$ (say), demonstrates that solutions for different parameters may pass through a common set of values. If the data correspond to the points of intersection, $\mathbf{p}$ is not uniquely determined. Even when there exists a unique $\widehat{\mathbf{p}}$, the success of iterative methods for determining its value may depend upon a sufficiently close starting approximation.

Identifiability of parameters, i.e., the uniqueness of the parameter minimizing the objective function $\Phi(\mathbf{p})$ is governed by the "shape" of the objective function in terms of the parameters. A visual assessment can be obtained from graphical displays of $\Phi(\mathbf{p})$ for a particular model (see FIG. I-2.2 and Part III [12] for examples). Although we do not address this question in depth, we regard this as an important procedure for providing practical insight, though it is complicated in the case of many parameters.

As illustrated in Figure I-2.2 (obtained from experimental data), it is instructive to obtain graphical illustrations of the behaviour of $\Phi(\mathbf{p})$ in neighbourhoods of $\widehat{\mathbf{p}}$. Valley-type behaviour of $\Phi(\mathbf{p})$ with a level valley floor (e.g., Figure I-2.2a) implies a high correlation between corresponding parameters: any combination of parameter components corresponding to the valley floor is equally effective at providing a small objective function. This interpretation can be revealed by analysis of the covariance matrix for parameter estimates $\boldsymbol{\Xi}(\mathbf{p})$ in (I-4.3).

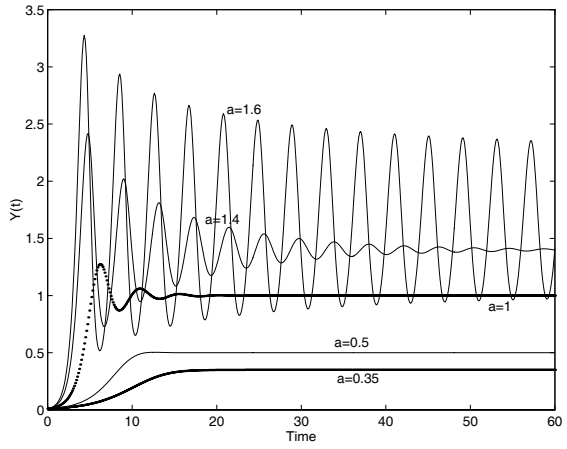

Figure I-2.1: Solutions of equation (I-2.4) with different parameters may intersect; data given at a countable set of points of intersection does not necessarily determine the parameters ($a = 1, 1.4.1.6$) uniquely.

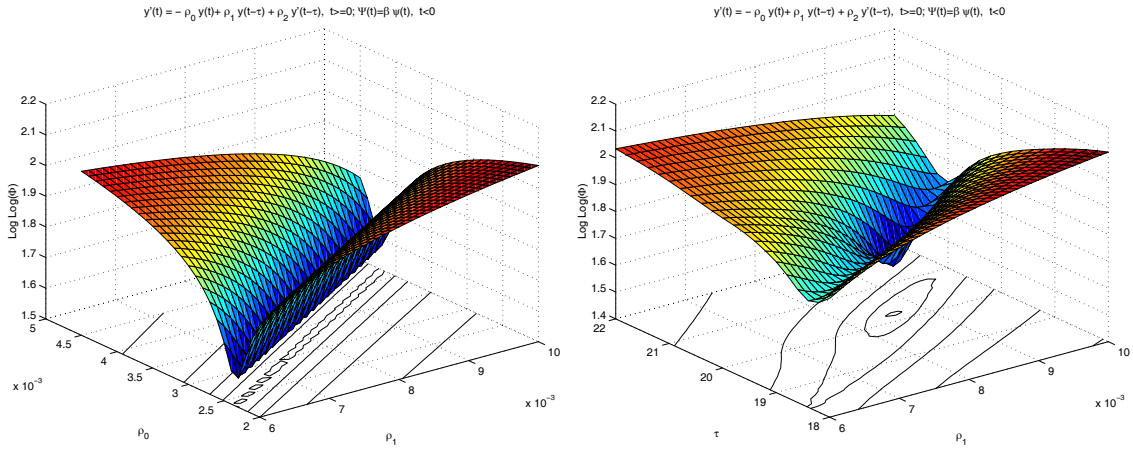

Figure I-2.2: Objective functions (with contour plots) displaying (a) valley-type behaviour and (b) locally elliptic contours, in the neighbourhood of a local minimum. In case (a) the minimum is ill-defined. The contours indicate the correlation of the parameters with each other. When the contours are close to a set of circles, this indicates a small correlation. General ellipticity (in a certain region of parameter space) indicates, at least locally, a small degree of nonlinearity of the model. For a discussion of methods of assessing parameter-effect nonlinearity and intrinsic nonlinearity, we refer to [50, Chapter 2]. The sensitivity coefficients provide a basis for assessing the nonlinearity of a model. Thus, a model is linear if $\frac{\partial}{\partial \mathbf{p}}\mathbf{y}(t;\mathbf{p})$ is a constant vector, and may be considered almost linear if this vector is almost constant.

The presence of errors in data sets affects the estimated parameters. It is desirable to specify the remaining uncertainty (e.g. 95% confidence intervals) in the estimates by quantifying the effect of the errors in data on the parameter estimates. Therefore, practical identifiability analysis involves examining, typically from the covariance matrix (see subsection I-4.3), the achievable accuracy of the parameter estimates obtained from the given noisy experimental data. To calculate the quantitative uncertainty in parameters a number of approaches exist, and these will be pursued in Parts II & III [11, 12].

I-3 Estimation of Model Parameters & Model Evaluation

We propose to discuss the selection of parameters in models and methods for the evaluation of resulting parametrized models. The approach to the latter question is based upon maximum likelihood. There is an interaction between the two issues, since the method for computing parameters depends upon an assessment of properties of the data and the model.

It is important, in our view, that the mathematical model should be consistent with the known science of the phenomenon being modelled. The concepts set forth in Part I will be applied to models of the type (I-2.1) in Parts II and III [11, 12], where they will be illustrated with reference to (I-2.1) and published observational data related to cell population dynamics. DDEs and NDDEs have richer dynamics than ODEs and in consequence they provide potentially more flexible tools for mathematical modelling. The work reported in [21] indicates the scope for applications of functional differential equations in bioscience. Our discussion adapts to scenarios in which the parametrized models are VIDEs, PDEs, etc.

Remark I-3.1 The complexity of DDEs and NDDEs requires us to consider possible discontinuities in the derivatives of the solution of the model. Such discontinuities are rarer when modelling with ODEs but discontinuities can arise for ODEs with non-smooth coefficients or forcing terms, and for certain Volterra integro-differential equations, as well as for parabolic and hyperbolic PDEs.

I-3.1 Objective functions for parameter estimation

A key element in defining a 'near-best' fit to data is the selection of an objective function. Practical features of the model or data sometimes influence the choice of the objective function. Different fitting criteria can be used [17] to reflect the (stochastic) features of the errors in the data. Given data $\{t_j; y_j^i\}_{j=1}^N$ (for $1 \le i \le M$) we choose a suitable[6] objective function $\Phi(\cdot)$, which depends upon the data and the values $\{y^i(t_j; \widehat{\mathbf{p}})\}_{j=1:N}^{i=1:M}$ of the parametrized solution $\mathbf{y}(t; \mathbf{p})$ of (I-2.1). We may seek the parameter $\widehat{\mathbf{p}}$ for (I-2.1) for which the corresponding values $\{y^i(t_j; \widehat{\mathbf{p}})\}_{j=1:N}^{i=1:M}$, provide a 'best fit' to the given $\{y_j^i\}_{j=1:N}^{i=1:M}$ in the sense that:

$$\Phi(\widehat{\mathbf{p}}) = \min_{\mathbf{p}} \Phi(\mathbf{p}). \tag{I-3.1}$$

We have $\Phi(\mathbf{p}) \ge 0$ and the size of $\Phi(\mathbf{p})$ (or more appropriately, its square root) is in general a measure of the corresponding *residual*, that is, a measure of the amount by which the values $\mathbf{y}(t_j; \mathbf{p})$ differ from the values $\mathbf{y}_j$.

Recall that we define the *gradient* of a smooth scalar-valued function of the column vector $\mathbf{q} \in \mathbb{R}^{k\times 1}$ as

$$\frac{\partial}{\partial \mathbf{q}} a(t; \mathbf{q}) := \Big[\frac{\partial}{\partial q_1} a(t; \mathbf{q}), \frac{\partial}{\partial q_2} a(t; \mathbf{q}), \cdots, \frac{\partial}{\partial q_k} a(t; \mathbf{q})\Big]^{\mathrm{T}} \in \mathbb{R}^{k\times 1}. \tag{I-3.2}$$

[6] In general, $\Phi(\cdot)$ is expressible as a seminorm of the error in fitting data by the solution of the model.

The corresponding row vector is written as[7]

$$\Big\{\frac{\partial}{\partial \mathbf{q}}\Big\}^{\mathrm{T}} a(t;\mathbf{q}) := \Big[\frac{\partial}{\partial q_1}a(t,\mathbf{q}), \frac{\partial}{\partial q_2}a(t;\mathbf{q}), \cdots, \frac{\partial}{\partial q_k}a(t;\mathbf{q})\Big] \ \in \mathbb{R}^{1\times k}. \tag{I-3.4}$$

Thus, the *gradient* of a smooth objective function $\Phi(\mathbf{p})$ is the vector $\dfrac{\partial}{\partial \mathbf{q}}\Phi(\mathbf{p})$ and the *Hessian matrix* is

$$\mathbf{H}(\mathbf{p}) \equiv \Big\{\frac{\partial^2}{\partial \mathbf{p}\partial \mathbf{p}^{\mathrm{T}}}\Big\}\Phi(\mathbf{p}) \quad := \quad \Big\{\frac{\partial}{\partial \mathbf{p}}\Big\}\Big\{\frac{\partial}{\partial \mathbf{p}}\Big\}^{\mathrm{T}}\Phi(\mathbf{p}) \in \mathbb{R}^{L\times L} \quad (\text{where } \mathbf{p}\in\mathbb{R}^L), \tag{I-3.5a}$$

assuming that the derivatives exist. In scalar notation, the Hessian is the matrix with (i,j)-th element $\dfrac{\partial^2}{\partial p_i \partial p_j}\Phi(\mathbf{p})$

Theorem I-3.1 *If the smooth function $\Phi(\mathbf{p})$ attains an unconstrained local minimum at $\widehat{\mathbf{p}}$ then the gradient vanishes at $\widehat{\mathbf{p}}$ and the Hessian is positive definite at $\widehat{\mathbf{p}}$. A non-smooth function $\Phi(\mathbf{p})$ assumes a constrained minimum either on a constraint or at a point where $\Phi(\mathbf{p})$ or its gradient is discontinuous or at a point where the gradient vanishes and the Hessian is positive semi-definite.*

Remark I-3.2 If in the previous statement the Hessian is positive-semi-definite, the minimum is not necessarily strict. To say that the Hessian $\mathbf{H}(\widehat{\mathbf{p}})$ at $\widehat{\mathbf{p}}$ is *positive-definite* (respectively, positive semi-definite) is to say that $\mathbf{H}(\widehat{\mathbf{p}}) \equiv \mathbf{H}(\widehat{\mathbf{p}},\mathbf{y}) := \dfrac{\partial^2}{\partial \mathbf{p}\partial \mathbf{p}^{\mathrm{T}}}\Phi(\widehat{\mathbf{p}})$ *has positive eigenvalues* (respectively, non-negative eigenvalues). Note that a positive-definite matrix $\mathbf{H}$ has diagonal elements h_{ii} that are positive.

Remark I-3.3 The objective function that we employ may not depend in a smooth manner upon τ; this has some implications if one seeks to apply standard computational techniques that are founded on the assumption that the objective function is a smooth function of the parameter values. For further remarks on differentiability with respect to parameters, see Baker & Paul [14], and Hartung & Turi [35].

Usually, the solution $\mathbf{y}(t;\mathbf{p})$ is *not* linear in all its parameters, and the objective function is in general not a quadratic function of the parameter. We then employ numerical algorithms for minimizing objective functions which are usually iterative procedures for computing parameter estimates and they require initial starting values. An obvious difficulty is that there is the possibility of the iterative scheme converging to a local minimum, or not converging at all, rather than converging to the desired global minimum. In general, computed evidence cannot provide guarantees, but (in our view) computation of contour plots is an invaluable aid to estimating whether one has a *global* minimum (and to assessing whether any global minimum is non-unique).

As a practical detail, the authors have noted elsewhere that the minimum $\widehat{\mathbf{p}}$ of a generic objective function $\Phi_\star(\{t_\ell\}_0^N;\mathbf{p})$ can often be approximated by the parameter $\widehat{\mathbf{p}}_n$ that minimizes $\Phi_\star(\{t_\ell\}_0^n;\mathbf{p})$

$$\begin{aligned}\Phi_\star(\{t_\ell\}_0^1;\mathbf{p}) \le \Phi_\star(\{t_\ell\}_0^2;\mathbf{p}) \le \cdots \\ \le \Phi_\star(\{t_\ell\}_0^n;\mathbf{p}) \le \Phi_\star(\{t_\ell\}_0^{n+1};\mathbf{p}) \quad \le \cdots \Phi_\star(\{t_\ell\}_0^N;\mathbf{p})\end{aligned} \tag{I-3.6a}$$

(nondecreasing) and

$$\widehat{\mathbf{p}}_n \approx \widehat{\mathbf{p}}_N =: \widehat{\mathbf{p}} \quad \text{for } n \text{ close to } N, \tag{I-3.6b}$$

Then $\widehat{\mathbf{p}}_n$ can be taken as a starting approximation in an iterative technique for determining $\widehat{\mathbf{p}}_{n+1}$ $(n=0,1,\ldots,N-1)$. We intend to consider relations like (I-3.6) in another place.

[7] For a vector $\mathbf{a}(t;\mathbf{q})\in\mathbb{R}^{k_1}$ where $\mathbf{q}\in\mathbb{R}^{k_2}$, we define the matrix

$$\Big\{\frac{\partial}{\partial \mathbf{q}}\Big\}^{\mathrm{T}} \mathbf{a}(t;\mathbf{q}) := \Big[\frac{\partial}{\partial q_1}\mathbf{a}(t;\mathbf{q}), \frac{\partial}{\partial q_2}\mathbf{a}(t;\mathbf{q}), \cdots, \frac{\partial}{\partial q_{k_2}}\mathbf{a}(t;\mathbf{q})\Big] \in \mathbb{R}^{k_1\times k_2}. \tag{I-3.3}$$

One alternative to increasing n toward N is to choose $\ell < L$, and fix $L - \ell$ of the parameter values $p_1, p_2, \ldots, p_L$ in order to determine the corresponding minimum, increasing ℓ toward L. This corresponds to examining the ℓ-th model in a nested sequence of models.

As a supplementary remark, one may be inclined to anticipate that reduction of the objective function to zero could be achieved simply by taking L to be sufficiently large. Basing a philosophy on this is not adequate; the scientific basis for the model and its parsimony must enter the discussion.

I-3.2 Least-squares and related objective functions

In our discussion of (I-3.1), we first focus on ordinary least squares (OLS) and weighted least squares (WLS) fitting. This corresponds respectively to objective functions $\Phi_{OLS}(\{t_\ell\}_0^N; \mathbf{p}) \equiv \Phi_{OLS}(\mathbf{p})$, $\Phi_{\mathcal{W}LS}(\{t_\ell\}_0^N; \mathbf{p}) \equiv \Phi_{\mathcal{W}LS}(\mathbf{p})$, where

$$\Phi_{OLS}(\mathbf{p}) := \sum_{j=1}^{N} \sum_{i=1}^{M} \left[y^i(t_j; \mathbf{p}) - y_j^i \right]^2 := \sum_{j=1}^{N} \| \mathbf{y}(t_j, \mathbf{p}) - \mathbf{y}_j \|^2; \tag{I-3.7a}$$

$$\Phi_{\mathcal{W}LS}(\mathbf{p}) = \sum_{j=1}^{N} \sum_{i=1}^{M} w_i^j \left[y^i(t_j, \mathbf{p}) - y_j^i \right]^2. \tag{I-3.7b}$$

We also consider weighted log-least-squares (WLogLS) fitting. This corresponds to an objective function $\Phi_{\mathcal{W}LogLS}(\{t_\ell\}_0^N; \mathbf{p}) \equiv \Phi_{\mathcal{W}LogLS}(\mathbf{p})$ where

$$\Phi_{\mathcal{W}LogLS}(\mathbf{p}) = \sum_{j=1}^{N} \sum_{i=1}^{M} w_i^j \left[\ \ell n(y^i(t_j, \mathbf{p})) - \ \ell n(y_j^i) \right]^2 \tag{I-3.7c}$$

where we denote the natural logarithm by $\ell n(\cdot)$. To use (I-3.7c), it will be assumed that $y_j^i > 0$ and $y^i(t_j; \mathbf{p}) > 0$. The objective function (I-3.7c) may be interpreted as a realization of (I-3.7b) in which the data are regarded as the values $\ell n(\mathbf{y}_j)$ and we consider a model (not necessarily of standard type) that has a solution $\ell n(\mathbf{y}(t_j; \mathbf{p})$; thus, in the case of (I-3.7c), any assumptions about the data must be valid for the values $\ell n(\mathbf{y}_j)$. In this context, important observations are given by Burnham & Anderson [25, p.81], which amounts (in essence) for the need for consistency in one's approach. Various objective functions correspond [32] to
(i) an assumption of arithmetic normality of observational errors (in which equivalent positive and negative deviations from expected values differ by equal amounts, and the objective function has the form (I-3.7a)) or
(ii) an assumption of geometric normality of observational errors (in which equivalent deviations differ by equal *proportions*). The latter assumption corresponds to a choice of (weighted) log-least-squares objective function (I-3.7c).

Having regard to the preceding observations, the notation $\Phi_\star(\mathbf{p})$ will be taken to denote a generic objective function which can be any one of (I-3.7a) – (I-3.7c).

Remark I-3.4 When the model is of the form (I-2.1) and τ is to be estimated, τ is a component of $\mathbf{p}$ that is constrained to be positive, and (I-3.1) is clearly a constrained minimization problem. Otherwise, the problem (I-3.1) is in principle an unconstrained minimization; however, in practice it is wise to introduce well-founded constraints on admissible parameters. In cell growth, the model with the parameter value $\mathbf{p}^\star$ cannot be regarded as realistic if the predicted populations are found to be negative (nor if the observations are negative!).

We return to a discussion of the choice of objective function. The objective function $\Phi_{\mathcal{W}LS}(\mathbf{p})$ can be expressed

$$\sum_{j=1}^{N} \{ [\mathbf{y}(t_j; \mathbf{p}) - \mathbf{y}_j]^T \mathbf{W}_j [\mathbf{y}(t_j; \mathbf{p}) - \mathbf{y}_j] \} \equiv \sum_{j=1}^{N} \| \mathbf{W}_j^{\frac{1}{2}} [\mathbf{y}(t_j, \mathbf{p}) - \mathbf{y}_j] \|^2, \tag{I-3.8a}$$

on defining a collection $\mathcal{W} = \{\mathbf{W}_j\}$ of matrices that we usually suppose are diagonal:

$$\mathbf{W}_j = \mathrm{diag}\{w_1^j, w_2^j, \ldots, w_M^j\}, \tag{I-3.8b}$$

where w_i^j are positive weights. The choice of the values $w_i^j > 0$ can be based on knowledge of the relative accuracy of the observed values $\mathbf{y}_j$.

In principle, there is nothing to prevent the use of a collection $\mathcal{W} = \{\mathbf{W}_j\}$ where each matrix $\mathbf{W}_j$ is positive-definite to define $\sum_{j=1}^N \|\mathbf{W}_j^{\frac{1}{2}}[\mathbf{y}(t_j;\mathbf{p}) - \mathbf{y}_j]\|^2$ as an objective function generalizing (I-3.7b). We shall concentrate on positive-definite diagonal matrices. It is often recommended that the weights be determined by the inverses of the variance of the error in the relevant data (see below). We return to details later, but, in the case

$$\sigma_j^2 \text{ is the variance } \mathcal{E}(\{\mathcal{E}(y_j^i) - y_j^i\}^2) \text{ (assumed constant for all } i),$$

where $\mathcal{E}(\cdot)$ denotes the expected value, we can propose

$$\mathbf{W}_j = \sigma_j^{-2}\mathbf{I}; \tag{I-3.8c}$$

in the case $\sigma_j = \sigma$ for all j, this becomes $\mathbf{W}_j = \sigma^{-2}\mathbf{I}$. As a further example of (I-3.8b) we may propose $w_i^j = \{\sigma y_j^i\}^{-2}$ and arrive at

$$\mathbf{W}_j = \sigma^{-2}\mathrm{diag}^{-2}[\mathbf{y}_j^T]. \tag{I-3.8d}$$

Here $\mathbf{I}$ is the identity matrix, and $\mathrm{diag}(\mathbf{z}^{\mathrm{T}}) = \mathrm{diag}([z_1, z_2, \ldots, z_m])$ denotes the diagonal matrix of order m with diagonal elements $z_1, z_2, \ldots, z_m$, so that $\mathrm{diag}^{-2}(\mathbf{z}^{\mathrm{T}}) = \mathrm{diag}([z_1^{-2}, z_2^{-2}, \ldots, z_m^{-2}])$.

For log-least-squares, we require some further notation. Given $\mathbf{z} = [z_1, z_2, \ldots, z_m]^T \in \mathbb{R}^m$ we employ the notation $\ell n(\mathbf{z})$, when $\mathbf{z}$ has positive components, to denote the vector

$$\ell n(\mathbf{z}) = [\ell n(z_1), \ell n(z_2), \ldots, \ell n(z_m)]^T \quad (\mathbf{z} > \mathbf{0})$$

and we recall that we write $\mathrm{diag}(\mathbf{z}^{\mathrm{T}}) = \mathrm{diag}([z_1, z_2, \ldots, z_m])$ to denote the diagonal matrix of order m with diagonal elements $z_1, z_2, \ldots, z_m$. We can then take for $\Phi_{\mathcal{W}LogLS}(\mathbf{p})$ the expression

$$\sum_{j=1}^N \{[\ell n(\mathbf{y}(t_j;\mathbf{p})) - \ell n(\mathbf{y}_j)]^T \mathbf{W}_j [\ell n(\mathbf{y}(t_j;\mathbf{p})) - \ell n(\mathbf{y}_j)]\} \equiv \sum_{j=1}^N \|\mathbf{W}_j^{\frac{1}{2}}[\ell n(\mathbf{y}(t_j,\mathbf{p})) - \ell n(\mathbf{y}_j)]\|^2, \tag{I-3.9a}$$

where we may propose

$$\mathbf{W}_j = \sigma^{-2}\mathbf{I}, \quad \text{or} \quad \mathbf{W}_j = \sigma_j^{-2}\mathbf{I}, \quad \text{or (in particular)} \ \mathbf{W}_j = \sigma^{-2}\mathrm{diag}^{-2}[\ell n(\mathbf{y}_j)]. \tag{I-3.9b}$$

Remark I-3.5 If we have two collections of (positive definite) weighting matrices $\mathcal{W}' = \{\mathbf{W}_j'\}$, and $\mathcal{W}'' = \{\mathbf{W}_j''\}$ where $\mathbf{W}_j' = s^2\mathbf{W}_j''$ for some non-zero s; then the minimizer $\widehat{\mathbf{p}}$ of $\Phi_{\mathcal{W}'LS}(\mathbf{p})$ is clearly the same as the minimizer of $\Phi_{\mathcal{W}''LS}(\mathbf{p})$ and from this viewpoint the factor s^2 is irrelevant. However, it is relevant in terms of the size of the objective functions at the minimimum, and, for an appropriate choice of weights $\mathcal{W}$, the minimum $\Phi_{\mathcal{W}LS}(\mathbf{p})$ has a rôle in the evaluation of certain indicators of the acceptability of the model.

I-3.3 Model evaluation based on an information-theoretic approach

The objective functions in §I-3.1 suggest a classical approach, but the outcome of a minimization of Φ is dependent on the 'correct' choice of objective function.

When one has confidence in the form of the model, the goodness of fit associated with parameter estimates $\widetilde{\mathbf{p}}$ can be characterized by the size of an objective function $\Phi_\star(\widetilde{\mathbf{p}})$. This is the data-fitting approach, and here $\widetilde{\mathbf{p}}$ may be an approximation (however obtained) to $\widehat{\mathbf{p}}$ satisfying $\Phi_\star(\widehat{\mathbf{p}}) = \min_{\mathbf{p}} \Phi_\star(\mathbf{p})$. Thus, one criterion by which to judge a model may be the size of $\Phi_\star(\widetilde{\mathbf{p}})$

[22]. However, if there is a *number* of candidate models, our task is not simply to identify one with the smallest objective function but to incorporate other criteria for discriminating between models of differing complexity. As observed by Myung [45], whereas least squares estimation in terms of an objective function $\Phi_\star(\mathbf{p})$ might be useful for obtaining a descriptive measure for the purpose of summarizing observed data, for statistical inference, such as model comparison, a more suitable approach is based upon maximum likelihood estimation. (We note that other approaches such as the Bayesian approach can be of value, but we shall not discuss these here.) As indicated above, an additional relevant criterion is that of model parsimony (simplicity of the model – the conservation of parameters or of complexity). Finally, an important criterion is consistency with *a priori* scientific theories (or the acceptance of consistent *a posteriori* theories); this attribute is difficult or impossible to formalize.

There are statistical criteria (such as the Akaike, Schwarz, and Takeuchi Information Criterion and generalizations of these related to informational complexity of models), which depend not only upon estimates of the maximum likelihood estimator (MLE) [2, 6, 53]. but incorporate the number of parameters and the number of observations, for a quantitative evaluation of different models. Bozdogan's criterion [23] is based upon a linear combination of the lack of fit (characterized by the size of an objective function), the lack of parsimony, and correlations between the parameter estimates; as Bozdogan [23] states

> "Other things being equal, the best model is the one which achieves the most satisfactory compromise between the accuracy of the estimated model parameters and the interaction of the residuals. The general principle is, for a given level of accuracy, a simpler model (i.e., the one with a small covariance matrix of the parameter estimates and a small residual covariance matrix) is preferable to a more complex one. Here small is used in the sense of minimum variance."

The practical effect of formal criteria such as those referred to above is, in brief, to moderate reliance upon the size of an objective function as an indicator of best fit. There exists a substantial literature on such criteria; we draw particular attention to the clear exposition presented by Burnham and Anderson [24, Chapters 2 & 6], [25, Chapters 2 & 7].

I-3.4 Indicators for distinguishing between models

The Akaike Information Criterion (AIC) [24, §2.2], [25, §2.2] and the corrected Akaike Information Criterion (c-AIC) [24, §2.4],[25, §2.4] (as well as other criteria such as Takeuchi's Information Criterion [24, p. 67],[25, p. 65] or the Schwarz Information Criterion "SIC" [24, p. 68], which we do not emphasize here) produce "indicators" that have been widely used as criteria for model selection in some areas of life-sciences [41, 42, 46, 60, 63]. The F test can also be used as a criterion for testing the goodness of fit of competing models; for a comparison of the F test with the Akaike criterion see [41, 42].

We suppose $\mathcal{L}(\widehat{\mathbf{p}})$ is the maximized likelihood, L is the number of parameters, M is the dimensionality of the state vector, and n is the sample size (the number of scalar observations); in the present discussion $n = NM$. In the case of the Akaike and the corrected Akaike criteria the indicators expressed in terms of the MLE are, respectively, the size of the measures μ_{AIC} and μ_{cAIC} given

$$\mu_{AIC} = -2\ \ell n\ \mathcal{L}(\widehat{\mathbf{p}}) + 2(L+1), \tag{I-3.10a}$$

$$\mu_{cAIC} = -2\ \ell n\ \mathcal{L}(\widehat{\mathbf{p}}) + 2(L+1) + \frac{2(L+1)(L+2)}{n-L-2}, \text{ with } n = NM, \tag{I-3.10b}$$

respectively; see [24, 25]. (The number of parameters being estimated is $L+1$, comprising $p_1, p_2, \ldots, p_L$ and σ, since we currently assume that a single value σ, which we also estimate, characterises all the variances.) The advice quoted in [24, p. 322],[25, p. 66] is that (I-3.10a) is satisfactory if $n < 40(L+1)$, otherwise (I-3.10b) is preferred by these authors. As $n \to \infty$, $\mu_{cAIC} \to \mu_{AIC}$. The quantities (I-3.10a) and (I-3.10b) differ by additive functions of L, and n and the literature contains discussion of the asymptotic correctness of the indicators (and the Schwarz indicator, for example) as $n \to \infty$; see [24, 25].

Remark I-3.6 In some scenarios, the data is suddivided into subsets, each of which is employed to estimate a subset of the parameters in the full model. Where the data is grouped, in this manner, into subsets (each of sample size n_i) the correction term in (I-3.10b) consists of a sum over i of terms in each of which n_i replaces n; see [24, p.255], [25, p.379].

The criterion of Akaike is related both to the maximum likelihood estimate (MLE) and to the Kullback-Leibler (K-L) notion of "distance" between two models (an information-theoretic concept). Burnham and Anderson [24] (see also [25]) review the concept of K-L information ("a dominant paradigm in information and coding theory") as a natural basis for model selection and maximum likelihood ("the dominant paradigm in statistics") and they describe the Akaike criterion, which Akaike related to the K-L theory in 1973, as "a new paradigm in statistical science".

Remark I-3.7 A number of papers relate to the indicators above. The books [24, 25] are useful references and provide bibliographies. The formulae (I-3.10) are special cases of the formulae

$$\mu_{AIC} = -2\ \ell n\ \mathcal{L}(\widehat{\mathbf{p}}) + 2\kappa, \tag{I-3.11a}$$

$$\mu_{cAIC} = -2\ \ell n\ \mathcal{L}(\widehat{\mathbf{p}}) + 2\kappa + \frac{2\kappa(\kappa+1)}{n-\kappa-1}, \tag{I-3.11b}$$

where n is the sample size and κ is the number of parameters being estimated (including the variances). Burnham and Anderson [24, 25] note that κ is often mistakenly to be the dimensionality of $\mathbf{p}$.

Cavanaugh [26] provided a unification of the derivations for the Akaike information criterion, AIC, and the corrected Akaike information criterion, c-AIC, which were both designed as estimators of the expected Kullback-Leibler discrepancy [39] between the model generating the data and a fitted candidate model. AIC is justified in a very general framework, and as a result, it appears to offer a crude estimator of the expected discrepancy: one which exhibits a potentially high degree of negative bias in small-sample applications. c-AIC corrects for this bias, but is less broadly applicable than AIC since its justification depends upon the form of the candidate model. Although AIC and c-AIC share the same objective, the original derivations of these criteria proceeded along different lines, making it difficult to reconcile how c-AIC improves upon the approximations leading to AIC. To address this issue, Cavanaugh presented a derivation which unifies the justifications of AIC and c-AIC in the linear regression framework.

In Section I-3.3, we shall relate (I-3.10) to the minimization of an objective function using weights that are determined by assumptions about the errors in the observations. The expressions defined in (I-3.10) are then replaced by expressions in terms of the minimum of an objective function.

Our aim is to analyze the family of models arising in cell growth dynamics, assisted by such criteria such as those mentioned above, and we supply examples later.

The additive terms in (I-3.24a)–(I-3.24b)), namely

$$2(L+1) \quad \text{and} \quad 2(L+1) + \frac{2(L+1)(L+2)}{n-L-2}, \quad \text{with} \quad n = NM \tag{I-3.12}$$

(modifying the size of the term involving $\mathcal{L}(\widehat{\mathbf{p}})$), can be regarded as "penalty functions" which have the *effect* of giving credit to parsimony (though they have their origins in other considerations). Burnham & Anderson ([24, §2.7.1]),[25, §2.12.2], rather than regarding the additive terms as penalty functions, prefer instead to motivate the additive terms from a consideration of Kullback-Liebler distance. The approach to this topic in [25] is somewhat modified in [25, pp.62–63].

The best model is often taken to be that which yields the lowest value of μ_{AIC} or μ_{cAIC} (or some similar criterion such as that of Schwarz or Takeuchi). As we have suggested, this happens to take account of parsimony, to some extent. However, the "penalty function" (just considered) often does not provide –at the least in the context of the models we discuss– a true measure of *computational complexity* of the model. The formal measurements of complexity normally found in the literature are governed by the number of parameters — which does not reflect the abundance or paucity of qualitative diversity. (A similar remark can be made in the comparison of nonlinear equations with linear equations, etc.)

We note that in modelling with DDEs, NDDE, or PDES, the *initial or boundary functions* defining a solution depend upon some or all of the parameters, whereas for a model governed by an ordinary differential equation the initial condition relates to an initial *value* – which often provides less rich dynamics. A delay differential equation with one parameter is effectively infinite-dimensional whereas an ordinary differential equation with one parameter is 1-dimensional. Since a DDE or NDDE with a fixed lag can (by the method of steps) be transferred to systems of ordinary differential equations of increasing dimensionality on successive intervals, there are functions of L, M, N which, for a fixed time-interval and a given lag, reflect the computational complexity compared with that of an ODE. What is apparent from this discussion is that, compared with the use of ODEs, there is a natural penalty to modelling with DDEs or NDDEs, just as there would be to modelling with PDEs: ODEs are simpler types of equations. However, what may be of practical interest is the time required to compute an approximate solution to a model of a specific type. Comparisons here depend upon the relative efficiency of the computational procedures.

I-3.5 Maximum likelihood approach

An appropriately chosen objective function $\Phi_\star(\widehat{\mathbf{p}})$ based on properly chosen weights can provide an approximation to the maximum likelihood estimator under certain assumptions. We shall indicate how one can proceed from an information theoretic criterion to a weighted least squares criterion under assumptions that we clarify in our discussion. Our systematic discussion is in part based on known results, but it extends to general M results that can be found for scalar equations (the case $M = 1$).

Criteria originating in information theoretic concepts, like those in (I-3.10a)) – (I-3.10b), that are based upon $\mathcal{L}(\widehat{\mathbf{p}})$ depend upon an assumption of the statistical nature of the errors in the observed data $\{\mathbf{y}_j\}_{j=1}^N$. If (Assumption I-3.5-1) the errors in the observed data are assumed to have a Gaussian distribution about the values $\{\mathbf{z}_j\}_{j=1}^N$, that is

$$\mathbf{y}_j \sim N(\mathbf{z}_j, \mathbf{\Sigma}_j),$$

where $\mathbf{\Sigma}_j$ is the j-th covariance matrix

$$\mathbf{\Sigma}_j := \begin{bmatrix} (\sigma_1^{[j]})^2 & \rho_{12}^{[j]}\sigma_1^{[j]}\sigma_2^{[j]} & \cdots & \rho_{1M}^{[j]}\sigma_1^{[j]}\sigma_M^{[j]} \\ \rho_{21}^{[j]}\sigma_2^{[j]}\sigma_1^{[j]} & (\sigma_2^{[j]})^2 & \cdots & \rho_{2M}^{[j]}\sigma_2^{[j]}\sigma_M^{[j]} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M1}^{[j]}\sigma_M^{[j]}\sigma_1^{[j]} & \rho_{M2}^{[j]}\sigma_M^{[j]}\sigma_2^{[j]} & \cdots & (\sigma_M^{[j]})^2 \end{bmatrix} \text{ with } \rho_{\ell k}^{[j]} = \frac{Cov(\mathbf{y}_j^\ell, \mathbf{y}_j^k)}{\sqrt{Cov(\mathbf{y}_j^\ell, \mathbf{y}_j^\ell) Cov(\mathbf{y}_j^k, \mathbf{y}_j^k)}} \tag{I-3.13}$$

($\rho_{\ell k}^{[j]}$ denoting the correlation coefficient of the ℓ-th and k-th observed component of $\mathbf{y}_j$) then the component probability density functions are given by

$$\left\{ \mathcal{H}(\mathbf{y}_j; \mathbf{p}) = \frac{1}{\sqrt{(2\pi)^M \det \mathbf{\Sigma}_j}} \exp\{-\frac{1}{2}[\mathbf{z}_j - \mathbf{y}_j]^T \mathbf{\Sigma}_j^{-1} [\mathbf{z}_j - \mathbf{y}_j]\} \right\}_{j=1}^N. \tag{I-3.14}$$

If (Assumption I-3.5-2) the errors in the components of $\mathbf{y}_j$ are assumed to be independent then $\rho_{\ell k}^{[j]} = 0$ (for $\ell \neq k$). In this case,

$$\mathbf{\Sigma}_j := \begin{bmatrix} (\sigma_1^{[j]})^2 & 0 & \cdots & 0 \\ 0 & (\sigma_2^{[j]})^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\sigma_M^{[j]})^2 \end{bmatrix} = \text{diag}([(\sigma_1^{[j]})^2, (\sigma_2^{[j]})^2, \cdots, (\sigma_M^{[j]})^2]). \tag{I-3.15}$$

When (Assumption I-3.5-3) the errors of observations at successive times are independent, the likelihood function is given by

$$\mathcal{L}(\mathbf{p}) = \prod_{j=1}^N \mathcal{H}(\mathbf{y}_j; \mathbf{p}), \tag{I-3.16}$$

where $\mathcal{H}(\mathbf{y}_j;\mathbf{p})$ is the probability density function (I-3.14). Thus, $\mathcal{L}(\mathbf{p})$ is a function of the matrices $\mathbf{\Sigma}_j$, but if the observation errors were *not* independent, the N matrices (I-3.14) would be inadequate to characterize the stochastic process. It would then be necessary to revise (I-3.16), but the changes are straightforward if $M = 1$ and in this case the correlations (*cf* [23]) enter into the amendments of our subsequent formulae.

Remark I-3.8 The probability density function (I-3.14) and the resulting likelihood function (I-3.16) correspond to a model assuming the normal distribution of the observations. In the case of a log-normal distribution the probability density function and, therefore, the likelihood, take a slightly different form (for details, see Burnham and Anderson [25, pp. 82 & 318].

If we now suppose (Assumption I-3.5-4) that $\mathbf{z}_j$ is the value $\mathbf{y}(t_j,\mathbf{p})$ then the expression $-2\ell n(\mathcal{L})$ required for use in equations (I-3.10a)–(I-3.10b) becomes, taking the natural logarithm of the product of the quantities in (I-3.14),

$$-2\ \ell n\ \mathcal{L}(\mathbf{p}) = NM\ell n(2\pi) + \sum_j \ell n(\det \mathbf{\Sigma}_j) + \sum_j \|\{\mathbf{\Sigma}_j^{-\frac{1}{2}}[\mathbf{y}(t_j;\mathbf{p}) - \mathbf{y}_j]\|^2. \tag{I-3.17}$$

From assumption (I-3.15),

$$-2\ \ell n\ \mathcal{L}(\mathbf{p}) = NM\ell n(2\pi) + 2\ell n(\prod_{i,j}\sigma_i^{[j]}) + \sum_j \|\{\mathbf{\Sigma}_j^{-\frac{1}{2}}[\mathbf{y}(t_j;\mathbf{p}) - \mathbf{y}_j]\|^2. \tag{I-3.18}$$

If the weights in $\Phi_{\mathcal{W}LS}(\mathbf{p})$ are chosen to match the variance of the error, so that, in (I-3.7b), $\mathbf{W}_j = \left\{\mathrm{diag}[\sigma_1^{[j]},\sigma_2^{[j]},\ldots,\sigma_M^{[j]}]\right\}^{-2}$, then the last term in (I-3.18) becomes $\Phi_{\mathcal{W}LS}(\mathbf{p})$. In particular, if $\mathbf{W}_j = \mathbf{\Sigma}_j^{-1}$ then (Assumption I-3.5-5)

$$\mathbf{W}_j = \sigma^{-2}\mathbf{\Omega}_j \text{ where } \mathbf{\Omega}_j = \left\{\mathrm{diag}[\omega_1^{[j]},\omega_2^{[j]},\ldots,\omega_M^{[j]}]\right\}^{-2}, \tag{I-3.19a}$$

(compare (I-3.8c)) then $\Phi_{\mathcal{W}LS}(\mathbf{p}) \equiv \Phi_{\mathcal{W}LS}(\mathbf{p},\sigma)$ and

$$\Phi_{\mathcal{W}LS}(\mathbf{p},\sigma) = \sigma^{-2}\Phi_{\Omega LS}(\mathbf{p}). \tag{I-3.19b}$$

This is to say $\Phi_{\mathcal{W}LS}(\mathbf{p},\sigma) \equiv \sigma^{-2}\sum_j \|\mathrm{diag}([\omega_1^{[j]},\omega_2^{[j]},\ldots,\omega_M^{[j]}])^{-1}[\mathbf{y}(t_j;\mathbf{p}) - \mathbf{y}_j]\|^2$. Then (I-3.18) yields

$$-2\ell n\ \mathcal{L}(\mathbf{p}) \equiv -2\ell n\ \mathcal{L}(\mathbf{p};\sigma) = NM\ell n(2\pi) + NM\ell n(\sigma^2) + 2\sum_{i,j}\ell n(\omega_i^{[j]}) + \sigma^{-2}\Phi_{\Omega LS}(\mathbf{p}) \text{ where } \Omega = \{\mathbf{\Omega}_j\}. \tag{I-3.19c}$$

The aim in the maximum likelihood strategy is to maximize (I-3.16), or equivalently (according to (I-3.19c)) to minimize (I-3.19b) by an appropriate choice of $\mathbf{p}$ and σ. Actually, the optimal value $\widehat{\sigma}$ of σ follows from the optimal value $\widehat{\mathbf{p}}$ of $\mathbf{p}$ and one may proceed by minimizing $\Phi_{\Omega LS}(\mathbf{p})$. Indeed, using the necessary condition,

$$\frac{\partial \ell n\ \mathcal{L}(\mathbf{p};\sigma)}{\partial \sigma^2} = 0,$$

for σ^2 to provide the maximum likelihood estimate we obtain the estimate

$$\widehat{\sigma}^2 = \frac{1}{NM}\sum_j \|\mathrm{diag}([\omega_1^{[j]},\omega_2^{[j]},\ldots,\omega_M^{[j]}])^{-1}[\mathbf{y}(t_j,\widehat{\mathbf{p}}) - \mathbf{y}_j]\|^2 = \frac{1}{NM}\Phi_{\Omega LS}(\widehat{\mathbf{p}}). \tag{I-3.20}$$

Remark I-3.9 Suppose Assumption I-3.5-5 is not valid, and $\mathbf{W}_j$ is independent of j

$$\mathbf{W}_j = \left\{\mathrm{diag}[\sigma_1,\sigma_2,\ldots,\sigma_M]\right\}^{-2}.$$

Then in place of (I-3.20) we estimate

$$\widehat{\sigma_i}^2 = \frac{1}{N}\sum_j |y^i(t_j;\widehat{\mathbf{p}}) - y_j^i|^2 \tag{I-3.21}$$

where $\widehat{\sigma_i}^2$ is the maximum likelihood estimator of the variance in the i-th component of the j-th observed vector (constant for all j). In this case, the total number of parameters estimated is $L+M$.

The quantity (I-3.18) (which is a multiple of the natural logarithm of the maximized likelihood) now becomes

$$-2\ell n\,\mathcal{L}(\widehat{\mathbf{p}};\sigma) = \left\{NM\ell n(2\pi)+NM+2\sum_{i,j}\ell n(\omega_i^{[j]})\right\}+NM\ell n(\Phi_{\Omega LS}(\widehat{\mathbf{p}}))-NM\ell n(NM). \tag{I-3.22}$$

Remark I-3.10 The standard least-squares estimator (see [24, p.17]) of σ^2 differs from (I-3.20) by a factor (which is immaterial in the present context).

The use of weights which are the elements of the inverse of the covariance matrix has not, to our knowledge, been justified for models with arbitrary data having non-normal distributions. Indeed, Himmelblau *et al.* [36] suggest that "observed dependent variables are taken over a successive sequence of times, they are not statistically independent as required for classical regression analysis". This is an area for future study. In the case where the errors of all observations are independent and of equal variance σ^2, $\mathbf{\Sigma}_j = \sigma^2\mathbf{I}$.

If the model is linear in the parameters (or if the number of observations is large) and the errors are normally distributed, then the choice of the weights (leading to least-variance estimates) is given by $\mathbf{W}_j = \Sigma_j^{-1}$, the elements of the inverse of the covariance matrix of the errors. Although we do not guarantee this optimal property in the general case (non-linear models with non-normal distributions), it is still common to use weights which are the elements of inverse of covariance matrix.

Theorem I-3.2 *In the case of normally distributed independent errors in the observations, equations (I-3.10a) and (I-3.10b) can be expressed in terms of* weighted least squares *estimation (using appropriate weights that reflect the variance of the errors in the observations) as:*

$$\mu_{AIC} = \left\{NM\ell n(2\pi)+NM+2\sum_{i,j}\ell n(\omega_i^{[j]})-NM\ell n(NM)\right\}+NM\ell n(\Phi_{\Omega LS}(\widehat{\mathbf{p}}))+2(L+1), \tag{I-3.23a}$$

$$\mu_{cAIC} = \mu_{AIC} + \frac{2(L+1)(L+2)}{NM-L-2}. \tag{I-3.23b}$$

Our interest is in the relative size of the indicators; thus, as long as the dimensionality of the state space and number of observations are fixed, it is convenient to discard extraneous terms and employ the revised indicators

$$\breve{\mu}_{AIC} = \frac{NM}{2}\ell n(\Phi_{\Omega LS}(\widehat{\mathbf{p}})) + 2(L+1), \tag{I-3.24a}$$

$$\breve{\mu}_{cAIC} = \breve{\mu}_{AIC} + \frac{2(L+1)(L+2)}{NM-L-2}. \tag{I-3.24b}$$

(The assumptions leading to the use of the parameters $\breve{\mu}_{\{.\}}$ are not valid in comparing a scalar equation with a coupled pair of equations, since the dimensionalities are not equal.) Then values $\breve{\mu}_{AIC}$ and $\breve{\mu}_{cAIC}$ that are assigned to the AIC and c-AIC criteria, depend on the number of parameter estimates, the total number of observations (which varies with the dimensionality of the state-space) as well as on the size of the objective function (the data-fit residuals) $\Phi_{\Omega LS}(\widehat{\mathbf{p}})$.

I-4 Feedback on the Structural Sensitivity of the Modelling Process

We here assess indications of sensitivity issues, bias, and the data sampling strategy. Each of these has an impact on the acceptability of a model or the process by which a model is arrived at. The sensitivity analysis has a multiple rôle: it can indicated redundancy of a parameter in the model, it can indicate predictability, but it can also indicate statistical properties of the data from which parameters are derived.

I-4.1 Sensitivity

Under the heading of sensitivity, we may discuss (a) The sensitivity of a solution $\mathbf{y}(t,\mathbf{p})$ to perturbations in $\mathbf{p}$ (notably, about $\mathbf{p}=\widehat{\mathbf{p}}$) (the sensitivity of the state variables to the parameters); (b) The sensitivity of the accepted estimator $\widehat{\mathbf{p}}$ to errors in the data; (c) The sensitivity of a solution $\mathbf{y}(t,\mathbf{p})$, at argument $t\in[t_0,T]$, to errors in the data. These are issues to which we return (in greater detail) in Part II [11], in the context of modelling with systems of NDDEs.

Remark I-4.1 Sensitivity analysis in modelling has been discussed elsewhere. It is appropriate here to cite Rabitz [49] who, in a different context, enumerates general applications of sensitivity analysis (without which a modelling process is "seriously deficient"), under the headings:

• Elementary (first-order) sensitivities	• Parameter–Observation interdependence (sensititivity to data)
• Parameter interdependence	• Interdependence of different observations
• Functional sensitivity analysis	• Higher-order sensitivities
• Memory effects and sensitivity analysis	• Position & time as dependent variables
• Sensitivity of objective functions	• Sensitivity to missing model components
• Structural sensitivity analysis	• Mapping out parameter space
	• Statistical error analysis.

If one considers practically significant perturbations, one may compute the differences (for example, $\mathbf{y}(t,\mathbf{p}+\delta\mathbf{p})-\mathbf{y}(t,\mathbf{p})$). If one considers the effect of 'infinitesimal' perturbations then one is led to an examination of derivatives (where they exist), involving sensitivity coefficients such as $\frac{\partial}{\partial p_i}\mathbf{y}(t,\mathbf{p})$. We shall indicate a use for such derivatives, below.

Note that the scaling of the parameter components p_i as they are represented in the model ((I-2.1), say) is a practical feature that may be important. The study of the *relative* sensitivity coefficients may prove helpful when scaling is an issue.

I-4.2 Nonlinearity and indications of bias

Bias of an estimator $\mathbf{p}$ is the difference between the expected value of the estimator and the true value, namely $b:=\mathcal{E}(\mathbf{p})-\mathbf{p}$; thus the estimator is *unbiased* if $b=0$. Nonlinear regression models differ from linear regression models in that, given the usual assumption of an independent and normally distributed errors, linear models give rise to unbiased, normally distributed, minimum variance estimators, whereas [50, p.13] nonlinear regression models have these properties only asymptotically (when the sample size becomes very large). Thus it is desired to estimate the impact or consequences of *nonlinearity* in the models. The bias in the parameters depends on the degree of nonlinearity of the structural model [41, 56].

The percentage bias in the parameter estimates serves as an indicator of the quantitative effect of nonlinearity [50]. Following Ratkowsky [50, §2.6], given a weighted least-squares parameter estimate $\widehat{\mathbf{p}}$, one may proceed as follows to examine its bias:

- Perturb the solution values $\mathbf{y}(t_i,\widehat{\mathbf{p}})$ of the model (corresponding to the best-fit parameters $\widehat{\mathbf{p}}$) by adding independent errors $\epsilon_j^i\sim N(0,\sigma^2)$ (see [17]),

$$\sigma^2=\frac{\Phi(\widehat{\mathbf{p}})}{n-\kappa},$$

  (wherein the total number of observations is n and κ is the total number of parameters);

- Find new best-fit parameter $\widetilde{\mathbf{p}}$ to the perturbed data $\left\{y^i(t_j,\widehat{\mathbf{p}})+\epsilon_j^i\right\}$ (where $1\le j\le N$ and $1\le i\le M$);

- Repeat this process sufficient times[8] to generate a statistically significant estimate of the mean value $\mathcal{E}(\widetilde{\mathbf{p}})$ of $\widetilde{\mathbf{p}}$.

If the *bias* satisfies the relation,

$$\|\widehat{\mathbf{p}} - \mathcal{E}(\widetilde{\mathbf{p}})\| < c\|\widehat{\mathbf{p}}\|, \text{ for a chosen (small) threshold value, } c \tag{I-4.1}$$

(that is, if the relative bias is below the chosen, small, threshold) then the bias of the weighted least-squares parameter estimate is considered not to be significant and the effect of non-linearity is deemed not to be significant; see Ratkowsky [50, Chapter 2]. One can then have confidence in the parameter estimates, and their standard deviations (see Part III [12]).

The importance of the question of linearity versus nonlinearity lies with the fact that the justification of various computations relies in places upon linear regression theory. For more details about nonlinearity effects in parameter estimations, we may refer to [27].

I-4.3 Improvement of the data sampling schedule

The mathematics should inform the design of experiment and, here, we consider the issue of a satisfactory choice of $\{t_j\}$ (there is a temptation to use equally-spaced values $\{t_j\}_{j=1}^N$ for no good reason). The observation interval could be divided into subintervals each of which could be informative about a specific parameter. Then, the sensitivity coefficients (which are functions of t) can be employed to assess qualitatively which data points have the most effect on a particular parameter; see [57].

As observed by Landaw and DiStefano [41], *ad hoc* sampling, particularly using equally spaced sampling points (which is conventional), can be inefficient in terms of the precision of the parameter estimates. Thus, the same precision can sometimes be obtained with fewer points, using D-optimal design. *With D-optimal design, we minimize* $\det \boldsymbol{\Xi}(\widehat{\mathbf{p}})$ *over the possible choices of* $\{t_j\}_{j=1}^N$*, where* $\boldsymbol{\Xi}(\widehat{\mathbf{p}})$ is the $L \times L$ covariance matrix for estimated parameters $\mathbf{p}$,

$$\boldsymbol{\Xi}(\mathbf{p}) \equiv \boldsymbol{\Xi}(\{t_j\}_1^N; \mathbf{p}) := \begin{bmatrix} \varsigma_{11} & \varsigma_{12} & \cdots & \varsigma_{1L} \\ \varsigma_{21} & \varsigma_{22} & \cdots & \varsigma_{2L} \\ \varsigma_{31} & \varsigma_{32} & \cdots & \varsigma_{3L} \\ \cdot\cdot & \cdot\cdot & \cdots & \cdot\cdot \\ \varsigma_{L1} & \varsigma_{L2} & \cdots & \varsigma_{LL} \end{bmatrix}, \quad \text{where } \varsigma_{rs} = Cov(p_r, p_s). \tag{I-4.2}$$

For $\boldsymbol{\Xi}(\widehat{\mathbf{p}})$ we have

$$\boldsymbol{\Xi}(\widehat{\mathbf{p}}) = 2\frac{\Phi(\widehat{\mathbf{p}})}{(NM - L)} \times [\mathbf{H}(\widehat{\mathbf{p}})]^{-1} \tag{I-4.3}$$

where $\mathbf{H}(\widehat{\mathbf{p}})$ is the Hessian matrix. The Hessian matrix can be approximated in terms of the "information matrix" of the objective function $\Phi_\star(\widehat{\mathbf{p}})$. Indeed, using the sensitivity coefficients

$$\mathbf{S}(t; \mathbf{p}) \equiv \left\{\frac{\partial}{\partial \mathbf{p}}\right\}^{\mathrm{T}} \mathbf{y}(t; \mathbf{p}) \in \mathbb{R}^{M \times L}. \tag{I-4.4}$$

we have

$$\mathbf{H}(\widehat{\mathbf{p}}) \approx 2\sum_{j=1}^{N} \mathbf{S}^T(t_j, \widehat{\mathbf{p}})\mathbf{W}_j\mathbf{S}(t_j, \widehat{\mathbf{p}}) =: \widetilde{\mathbf{H}}(\widehat{\mathbf{p}}). \tag{I-4.5}$$

(The matrix $\sum_{j=1}^N \mathbf{S}^T(t_j, \widehat{\mathbf{p}})\mathbf{W}_j\mathbf{S}(t_j, \widehat{\mathbf{p}})$ is the information matrix.) The above result holds for $\Phi_{OLS}(\mathbf{p})$, $\Phi_{\mathcal{W}LS}(\mathbf{p})$, generalizations to $\Phi_{WLogLS}(\mathbf{p})$ being possible. Further discussion appears in Part II [11]: We can calculate an exact value of $\det \boldsymbol{\Xi}(\widehat{\mathbf{p}})$ from the first and second order sensitivity coefficients, discussed in Part II [11].

[8] Ratkowsky [50] advises us to repeat the process 1000 times.

$\Xi(\mathbf{p})$ provides a measure of the dispersion of the best-fit estimate about its mean, and if the estimator is *unbiased* (i.e. the difference $\mathcal{E}(\widehat{\mathbf{p}}) - \widehat{\mathbf{p}}$ is small) then $\Xi(\widehat{\mathbf{p}})$ characterizes the spread of the multivariate distribution around $\widehat{\mathbf{p}}$.

Following the D-optimal design framework [30] the problem can be treated as follows: find the sample times that minimize the determinant of the covariance matrix (i.e. the volume of the asymptotic confidence region for $\widehat{\mathbf{p}}$). That is,

$$\text{determine}\{t_j^*\}_{j=1}^N \text{ to minimize } \det[\Xi(\{t_j\}_{j=1}^N, \widehat{\mathbf{p}})]. \tag{I-4.6}$$

The minimization may be performed for fixed N and the effect of varying N subject to practical constraints can be observed.

I-5 Computational Aspects

I-5.1 Numerical tools for differential equations

Evaluation of the objective function $\Phi(\mathbf{p})$ requires us to solve a DDE or a NDDE or a PDE with given parameters which we need to revise in an iterative manner in order to determine $\widehat{\mathbf{p}}$. An analytical approach, to computing $\Phi(\mathbf{p})$, is in general unrealistic and we use a numerical approach.

Reliable and robust software for solving ODEs and PDEs, as well as optimization, is available in most numerical software libraries, for example,

- the NAg library `http://www.nag.co.uk` and
- the IMSL `http://www.imsl.com`, NETLIB (`http://www.netlib.org`) and the HSL – formerly "Harwell Scientific Library" – `http://www.cse.clrc.ac.uk/nag/hsl`,
- MATLAB `http://www.mathworks.com`.

In the case of DDEs and NDDEs, although they have been the subject of much research, software production is still an active area of research [47]. (For NDDEs the question of the continuity, existence and uniqueness of solutions is more complicated than in the case of DDEs.) Links to various codes can be found at `http://www.maths.man.ac.uk/~chris/software`; we draw attention to [29, 34, 48, 55]. We refer to [47] for remarks on the design and analysis of numerical methods for DDEs and NDDEs. Our numerical work based upon DDEs and NDDEs has relied on one-step *continuous* RK methods. Some PDEs with delay can be solved by applying the method of lines to obtain a systems of DDEs that can then be solved by existing DDE methods. Codes for PDEs tend to be type-specific rather than general purpose. Attempts have been made to distinguish between "stiff" and "non-stiff" DDEs; a distinction which even for ODEs has been regarded as not robust (see [1]). We use the term "stiff" to indicate that step-size in the numerical integration of a particular solution is controlled by the demands of stability (or conditioning) rather than of local accuracy; in this case, stiffness can be transient or long-term and depends upon the particular solution.

I-5.2 Properties of objective functions

The task of parameter estimation is one of minimizing a suitable objective function based on the unknown parameters and observed data. In the case of parameter estimation for DDEs, this can include not only estimating parameters appearing in the DDEs but also estimating the position of the initial point, the initial function and the delayed arguments.

Most numerical software libraries include routines for minimizing an objective function, although these routines generally require that the objective function is at least continuous and more often has continuous first and even second order derivatives. Whilst these smoothness requirements are usually satisfied by ODE models, the same is not always true for DDE or NDDE based models [14]. DDEs and NDDEs exhibit a propagation of discontinuities and derivative discontinuities

in solutions [15, 31, 62], that must be handled correctly in order to develop efficient and robust software [47]. (Analogous difficulties can arise in the solution of partial differential equations, notably hyperbolic PDEs.) As a practical detail that is potentially of some importance, in certain (identifiable) circumstances [14], $\Phi_\star(\mathbf{p})$ may not possess the smoothness properties required by many techniques for minimizing $\Phi_\star(\mathbf{p})$. However, for the models that we consider here, it appears that this lack of smoothness in the objective function does not affect the correctness of the best-fit parameter values; we always checked the best-fit parameters values by examining contour plots of the objective function.

I-5.3 Minimization of objective functions

In the context of minimizing a given objective function $\Phi_*(\mathbf{p})$ the optimum parameter $\widehat{\mathbf{p}}$ is taken to be the value such that

$$\Phi_*(\widehat{\mathbf{p}}) \leq \Phi_*(\mathbf{p}) \quad \text{for all physically meaningful values of } \mathbf{p} \text{ and } \widehat{\mathbf{p}}.$$

Given a set of experimental data, the technique for finding the best-fit parameter values for a given mathematical model and objective function consists of the following steps:

Providing an initial guess $\mathbf{p}_0$ for the parameter estimates.

Solving the model equations using the current values of the parameters in order to compute $\Phi_*(\mathbf{p})$.

Adjusting the parameter values (by the minimization routine, for example `E04UPF` (succeeded by `E04USF`) in[9] the `NAG` library, `LMDIF` from[10] NETLIB and `FMINS` in MATLAB) so as to reduce the value of the objective function; see [48].

Minimization routines may be classed as "derivative-free" or "gradient-dependent". For gradient dependent methods, the gradient may be supplied analytically or computed numerically. (In the case of models of the type discussed here, the gradient can sometimes be computed by solving supplementary equations, at the cost of additional computational expense.)

When no further reduction in the value $\Phi_*(\mathbf{p})$ is possible, the best fit parameter values have been found. In order to find the *global* best-fit parameter values, the initial parameter values should be sufficiently close to the true minimum. Thus, good starting estimates for the parameter values can be of great assistance, both in speeding up the minimization process and finding the global minimum. Local minimum can also be avoided by repeating the iterative scheme for a variety of different initial estimates of parameter vector. It can be beneficial to display graphically the behavior of the function $\Phi_*(\mathbf{p})$; see FIG. I-2.2.

I-6 Conclusions

We have reviewed the principles governing the selection of a parametrized model chosen to give quantitative and qualitative agreement with processes giving rise to observed data, and the construction of an associated methodology. There is a balance to be achieved between agreement with the data (under-fitting and over-fitting) and the principle of parsimony. The theoretical foundations of what we propose rely upon assumptions – often made implicitly by researchers – concerning errors in the data (usually, that these are, for example, independently identically normally distributed). Estimation of the variances permits a rational choice of the weights in the weighted least-squares objective function.

Exploiting the link between a (weighted) least-squares fit, information theory, and maximum likelihood, we can employ indicators that can be regarded as taking parsimony into account. A

[9] `E04USF`, which replaced `E04UPF` at Mark 20 of the Library, is designed to minimize an arbitrary smooth sum of squares function subject to constraints (which may include simple bounds on the variables, linear constraints and smooth nonlinear constraints) using a sequential quadratic programming (SQP) method.

[10] `LMDIF` is an unconstrained minimization routine based on the Levenberg-Marquardt algorithm.

sensitivity analysis provides feedback on the covariances of the parameters in the model and on the need to include all the parameters possible. In a number of cases we have extended results in the literature in order to accommodate *systems* of differential equations as opposed to scalar equations. We have also remarked upon the impact of derivative discontinuities (such as those associated with the solutions of DDEs, NDDEs).

In subsequent Parts II & III [11, 12] we shall develop and apply the ideas presented here.

I-7 Bibliography for Part I

References

[1] Aiken, R.C. *Stiff Computation*, Oxford University Press, New York (1985), ISBN: 0-19-503453-8.

[2] Akaike H., A new look at the statistical model identification, *IEEE Trans. Automatic Control*, **19** (1974) 716-723.

[3] Anderson, D. H., *Compartmental Modeling and Tracer Kinetics*, Lecture Notes in Biomathematics, 50. Springer-Verlag, Berlin (1983), ISBN: 3-540-12303-2.

[4] Anderson, R.M., and May, R.M., *Infectious Diseases of Humans. Dynamics and Control*, Oxford University Press, Oxford (1991), ISBN: 0-19-854599-1.

[5] Appleton, D.R., An overview of models of cell proliferation, *Journal of Theoretical Medicine* **1** (1997) 53–62.

[6] Armitage P., Berry G., and Matthews J.N.S., *Statistical Methods in Medical Research*. (Fourth Edition) Blackwell Science, Oxford (2001), ISBN: 0-63-205257-0.

[7] Audoly, S., Bellu, G., D'Angio, L., Saccomni, M., and Cobelli, C., Global identifiability of nonlinear biological system, *IEEE Trans. Biomedical Eng.* **48** (2001) 55–65.

[8] Audoly, S., D'Angio, L., Saccomni, M., and Cobelli,C., Global identifiability of linear compartmental models – A computer algebra algorithm, *IEEE Trans. Biomedical Eng.* **45** (1998) 33–47.

[9] Bailey, N., *The Mathematical Theory of Infectious Diseases* (2nd ed.), Charles Griffin. London, (1975), ISBN 0-85264-231-8.

[10] Baker, C.T.H., Bocharov, G.A., Paul, C.A.H., Rihan, F.A., Modelling and analysis of time-lags in some basic patterns of cell proliferation, *J. Math. Biol.* **37** (1998) 341-371.

[11] Baker, C.T.H., Bocharov, G.A., Paul, C.A.H., Rihan, F.A., Models with Delays for Cell Population Dynamics: Identification, Selection and Analysis – Part II. In preparation.

[12] Baker, C.T.H., Bocharov, G.A., Paul, C.A.H., Rihan, F.A., Models with Delays for Cell Population Dynamics: Identification, Selection and Analysis – Part III. In preparation.

[13] Baker, C.T.H., and Parmuzin, E.I., Delay differential equations: identification of the initial function, *MCCM report, University of Manchester* (in preparation), ISSN 1360-1725.

[14] Baker, C.T.H., and Paul, C.A.H., Pitfalls in parameter estimation for delay differential equations, *SIAM J. Sci. Comp.* **18** (1997) 305–314.

[15] Baker, C.T.H. and Paul, C.A.H, Piecewise continuous solution of neutral delay differential equations, *MCCM report, University of Manchester* (in preparation), ISSN 1360-1725.

[16] Banks, R.B., *Growth and Diffusion Phenomena. Mathematical Frameworks and Applications*, Springer-Verlag, Berlin (1994), ISBN 3-540-55507-2.

[17] Bard, Y., *Nonlinear Parameter Estimation*, Academic Press, New York (1974).

[18] Bates, D.M., and Watts, D.G., *Nonlinear Regression Analysis and its Applications* John Wiley, New York (1988), ISBN 0-471-81643-4.

[19] Bellman, R., and Åström, K.M., On structural identifiability, *Math. Biosci.* **7** (1970) 329–339.

[20] Bocharov, G.A., and Hadeler, K.P., Structured population models, conservation laws, and delay equations, *J. Diff. Eqs.* **168** (2000) 212-237.

[21] Bocharov, G.A., and Rihan, F.A., Numerical modelling in biosciences using delay differential equations, *J. Comput. Appl. Math.* **125** (2000) 183-199.

[22] Borghans, J.A., Taams, L.S., Wauben, M.H.M., and De Boer, R.J., Competition for antigenic sites during T cell proliferation: A mathematical interpretation of *in vitro* data, *Proc. Natl. Acad. Sci. USA.* **96** (1999) 10782–10787.

[23] Bozdogan, H., Akaike's information criterion and recent developments in information complexity, *J. Math. Psych.* **44** (2000) 62–91.

[24] Burnham, K.P., and Anderson, D.R.: *Model Selection and Inference - a practical information-theoretic approach*, Springer-Verlag, New York (1998), ISBN 0-387-98504-2.

[25] Burnham, K.P., and Anderson, D.R.: *Model Selection and Multimodel Inference - a practical information-theoretic approach; 2nd ed.*, Springer-Verlag, New York (2002), ISBN 0-387-95364-7.

[26] Cavanaugh, J.E. Unifying the derivations for the Akaike and corrected Akaike information criteria, *Statist. Probab. Lett.* **33** (1997) 201–208.

[27] Chambers, J., and Hill, M., Fitting nonlinear models: numerical techniques, *Biometrika* **60** (1973) 1-13.

[28] Cobelli, C., and DiStefano, J., Parameter and structural identifiability concepts and ambiguities: a critical review and analysis, *Am. J. Physiol.* **239** (1980) R7-R24.

[29] Engelborghs, K., Internet links to software for delay differential equations `http://www.cs.kuleuven.ac.be/~koen/delay/software.shtml`

[30] Fedorov V.V., *Theory of Optimal Experiment Design*, Academic Press, New York (1972).

[31] Neves, K.W. and Feldstein, A., Characterization of jump discontinuities for state dependent delay differential equations, *J. Math. Anal.* **56**(1976) 689–707.

[32] Gingerich, P.D., Arithmetic or Geometric Normality of Biological Variation: an Empirical Test of Theory, *J. theor. Biol.* **204** (2000) 201–221.

[33] Gopalsamy, K. *Stability and oscillations in delay differential equations of population dynamics*, Kluwer Academic, Dordrecht (1992), ISBN 0-79-231594-4.

[34] Hairer, E., Codes for solving DEs. `http://www.unige.ch/math/folks/hairer/software.html` and `http://ftp.zib.de/elib/hairer-wanner/nonstiff/dr_retard.f`

[35] Hartung, F. and Turi, J., On differentiability of solutions with respect to parameters in state-dependent delay equations, *J. Diff. Eqns.* **135** (1997) 192–237.

[36] Himmelblau, D.M., Jones, C.R., and Bischoff, K.B., Determination of rate constants for complex kinetic models, *I & EC Fundamentals.* **6** (1967) 539–543.

[37] Hopkins, J.C., and Leipold, R.J., On the dangers of adjusting the parameter values of mechanism masked mathematical models, *J. theor. Biol.* **183** (1996) 417–427.

[38] Kolmanovskii V.B. and Myshkis A.D., *Applied theory of functional differential equations.* MIA vol. 85, Kluwer Academic, Dordrecht (1992).

[39] Kullback, S., Leibler R.A. On information and sufficiency, *Ann. of Math. Stat.*, **22** (1951) 79-86.

[40] Kuang, Y. *Delay Differential Equations with Applications in Population Dynamics*, Academic, Boston, (1993), ISBN 0-12-427610-5.

[41] Landaw, E.M., and DiStefano, J.J., Multiexponential, multicompartmental, and noncompartmental modeling. II Data analysis and statistical considerations, *Am. J. Physiol.* **5** 665–677.

[42] Ludden, T.M., Beal, S.L., and Sheiner, L.B., Comparison of the Akaike Information Criterion, the Schwarz Criterion and the F Test as Guides to Model Selection, *J. Pharmacokinetics & Biopharmaceutics* **22** (1994) 431–445.

[43] Marchuk, G. I., *Adjoint Equations and Analysis of Complex Systems* (translated from the 1992 Russian edition by G. Kontarev and revised by the author). MIA vol. 295, Kluwer Academic, Dordrecht, 1995, ISBN: 0-7923-3013-7.

[44] Marchuk, G. I., *Mathematical Modelling of Immune Response in Infectious Diseases* (translated from the Russian by G. Kontarev and I. Sidorov.) Kluwer Academic, Dordrecht, 1997, ISBN: 0-7923-4528-2.

[45] Myung, I.J., Maximum likelihood estimation, submitted for publication, 2001 (available at URL: `http://quantrm2.psy.ohio-state.edu/injae/respub.htm`).

[46] Myung, I.J., Forster, M. R., and Browne, M. W. (Guest Editors), Special Issue on Model Selection. *J. Math. Psych.* **44** (2000) 1-231.

[47] Paul, C.A.H., Designing efficient software for solving delay differential equations, *J. Comput. Appl.* Math. **125** (2000) 287–295.

[48] Paul, C.A.H., A user-guide to `Archi` - An explicit Runge-Kutta code for solving delay and neutral differential equations and parameter estimation problems, *MCCM report No. 283 (ISSN 1360-1725), University of Manchester* (1997) `http://www.ma.man.ac.uk/MCCM/MCCM.html`).

[49] Rabitz, H., Chemical sensitivity analysis theory with applications to molecular dynamics and kinetics, *Computers & Chemistry.* **5** (1981) 167–180.

[50] Ratkowsky, D.A., *Nonlinear Regression Modeling, A Unified Practical Approach* Marcel Dekker (1983), ISBN: 0-8247-1907-7

[51] Rubinow, S.I., *Mathematical Problems in the Biological Sciences*, SIAM Publication, Philadelphia, 1973.

[52] Sakamoto, T., Ishiguro, M., and Kitagawa, G, *Akaike Information Criterion Statistics*, D. Reidel, Holland (1986) (translated by the authors from the original Japanese, Kyoritsu Publishing Company (1983)), ISBN: 9-027-72253-6.

[53] Schwarz G., Estimating the dimension of a model, *Annals of Statistics* **6** (1978) 461-464.

[54] Sclove, S.L., Some aspects of model-selection criteria,*Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: an Informational Approach, Vol. 2 (Knoxville, TN, 1992)*, 1–5, 37–67, Kluwer Academic, Dordrecht (1994).

[55] Shampine, L.F., Thompson, S., and Kierzenka, J., Solving delay differential equations with dde23, (ftp://ftp.mathworks.com/pub/doc/papers/dde/)

[56] Sheiner, L.B., and Beal, S.L., Pharmcokinetic parameter estimates from several least squares procedures: superiority of extended least squares, *J. Pharmacokinetics & Biopharmaceutics* **13** (1985) 185–201.

[57] Thomaseth, K., and Conelli, C., Generalized sensitivity in physiological system identification, *Annals Biomed. Engin.* **27** (1999) 607–616.

[58] Tikhonov A. and Arsenin V., *Solution of Ill-posed Problems.* John Wiley, New York (1977).

[59] Verduyn Lunel, S.M., Parameter identifiability of differential delay equations, *Int. J. Adapt. Control Signal Process* **15** (2001) 655-678.

[60] Verotta, D., Schaedeli, F., Non-linear dynamics models characterizing long-term virological data from AIDS clinical trials, *Math. Biosci.* **176** (2002) 163-183.

[61] Voit, E.O., *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular biologists*, Cambridge University Press, Cambridge (2000), ISBN 0-521-78579-0.

[62] Willé, D.R. and Baker, C.T.H. The tracking of derivative discontinuities in systems of delay-differential equations, *Appl. Numer. Math.* **9** (1992) 209–222.

[63] Wolters, L.M.M., Hansen, B.E., Niesters, H.G.M., Levi-Drummer, R.S.L., Neumann, A.U., Schalm, S.W., de Man, R.A., The influence of baseline characteristics on viral dynamic parameters in chronic hepatitis B patients treated with lamivudine, *J. Hepatology* **37** (2002) 253-258.