Computational modelling with functional differential equations: Identification, selection, and sensitivity


Mathematics Department, Manchester University, Manchester M13 9PL, UK
Available online 14 October 2004

Abstract

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 (based on observational data) for the computational identification of, and discrimination between, quantitatively consistent models, using scientifically meaningful parameters.

We propose that a computational approach is essential for obtaining meaningful models. For example, it permits the choice of realistic models incorporating a time-lag which is entirely natural from the scientific perspective. The time-lag is a feature that can permit a close reconciliation between models incorporating computed parameter values and observations. Exploiting the link between information theory, maximum likelihood, and weighted least squares, and with distributional assumptions on the data errors, we may construct an appropriate objective function to be minimized computationally. The minimizer is sought over a set of parameters (which may include the time-lag) that define the model. Each evaluation of the objective function requires the computational solution of the parametrized equations defining the model. To select a parametrized model, from amongst a family or hierarchy of possible best-fit models, we are able to employ certain indicators based on information-theoretic criteria. We can

---

∗ Corresponding author.

E-mail addresses: chtbaker@ma.man.ac.uk, cthbaker@na-net.orl.gov (C.T.H. Baker), bocharov@inm.ras.ru (G.A. Bocharov), chris@ma.man.ac.uk, cpaul@na-net.orl.gov (C.A.H. Paul), frihan@salford.ac.uk, frihan@na-net.orl.gov (F.A. Rihan).

1 Research Professor, Manchester University; Visiting Professor, University College Chester.

2 INM, Russian Academy of Sciences, Moscow; Honorary Research Fellow, Manchester; Leverhulme Visiting Professor, University College Chester. Supported by the Leverhulme Trust and the Russian Foundation for Basic Research.

3 Research Fellow.

4 University of Salford; Honorary Research Fellow, Manchester. Supported in part by NERC and the Egyptian Government.
evaluate confidence intervals for the parameters, and a sensitivity analysis provides an expression for an information matrix, and feedback on the covariances of the parameters in relation to the best fit. This gives a firm basis for any simplification of the model (e.g., by omitting a parameter).

Keywords: Computation; Data; Differential equations; Identifiability; Information-theoretic criteria; Modelling; Objective function; Parametric estimation; Sensitivity; Time-lag; Well-posedness

1. Introduction

One of the significant challenges in bio-mathematics (and other areas of science) is to formulate meaningful mathematical models. For example, as indicated in [32], “The modern ecologist usually works in both the field and laboratory, uses statistics and computers, and often works with ecological concepts that are model-based, if not model-driven.”

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 moderately changed conditions. In this sense, a good model will reduce the need to perform certain experiments; however, the production of a good model depends on the availability of sufficient data of sufficient quality. We shall 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; our approach to model identification is related to (but has a philosophy different from) data assimilation; cf. [38].

Many of the aspects discussed below apply to the construction of models involving ordinary or partial differential equations (ODEs or PDEs), but in the area of interest the incorporation of time-lag is an essential feature of realistic models. In many studies in the literature, unrealistically simple models (whose structure is constrained by the need to find closed-form solutions or to succumb to theoretical analysis) appear to be employed. We believe that a computational approach is essential for obtaining meaningful models, and there is a need to recognize a practical role for numerical analysts since successful modelling depends on sound computational practice. Issues in the identification of mathematical models impact on the software that is required and ought to inform the numerical analysis community of areas of study that merit attention.

ODEs, PDEs, and integral equations (IEs) have all been given rôles in mathematical models of, e.g., biological, chemical, and physical phenomena, see [3,16]. Models incorporating hereditary effects (delay, neutral, integral or integro-differential equations) are discussed in [19,29,35,39]. Conceptually, it is a small step to extend ODE models to admit delay differential equations (DDEs) and neutral delay differential equations (NDDEs), which have a richer dynamical structure; these are, however, unlikely to have closed-form solutions. The preceding types of equation are all examples of differential equation or functional differential equation. Various types of model have their supporters, often irrespective of the context in which they are proposed. If one constructs a family or hierarchy of models (which need not be

\[^5\] Integro-differential equation formulations that involve continuously distributed time-lags can sometimes be approximated by DDEs with multiple time-lags.
a nested or well-ordered hierarchy), an approach to the selection of a particular model from the hierarchy is valuable.⁶

Many scientists have considered the difficulties associated with mathematical modelling. As Hopkins and Leipold [33] 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.”

We have an interest in modelling the growth dynamics of in vitro and in vivo biological systems (such as cell proliferation). However, the mathematical and computational strategies we employ have features to be found in other areas, and our remarks have widespread applicability. In [9,10], various levels of model complexity were provided by formulating a hierarchy of cell-growth models based on different classes of equations incorporating time-lags, each requiring the specification of scientifically meaningful parameters. Appleton, in a presentation on which [4] is based, listed fundamental questions related to parametrized models of cell proliferation (they are relevant to more general modelling problems). The list included: (i) How should the best fit be assessed? (ii) How should the model equations be solved? (iii) Which model is better: that having the fewer parameters or that based more closely on the biology of the system? (iv) Is it possible to distinguish between the important and the unimportant assumptions in a model?

Mathematical discussions found in the literature are often based on 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 to estimate parameters in models that should be both qualitatively and quantitatively consistent with actual observations. We regard soundly based and ‘robust’ numerical techniques as essential for quantitatively consistent modelling.

2. Fundamental issues concerning modelling

The general discussion here will be conducted with reference to models based on ODEs, DDEs, and NDDEs (cf. [15]). The computational approach permits the choice of realistic (and, if necessary, quite complex) equations. The ODEs, DDEs and NDDEs considered as potential models have solutions that we denote \( y(t) = y(t; p) \in \mathbb{R}^M \), with parameter vector \( p = [p_1, p_2, \ldots, p_L]^T \in \mathbb{R}^L \); the models often have the following general form:

\[
\begin{align*}
\dot{y}'(t; p) &= f(t, y(t; p), y(t - \tau; p), y'(t - \tau; p); p), \quad \text{for } t \in [t_{\text{start}}, T]; \\
y(t; p) &= \psi_0(t; p), \quad y'(t; p) = \psi_1(t; p), \quad \text{for } t \in [t_{\text{start}} - \tau, t_{\text{start}}].
\end{align*}
\] (2.1)

Components \( p_\ell \) (one of which may be \( \tau \)) of the parameter vector \( p \) will be called “parameters”; \( \tau \geq 0 \); if \( \tau > 0 \) it represents a time-lag. The methodology outlined here gives, inter alia, an indication whether a lag parameter \( \tau > 0 \) can be justified. The form of \( f \) is known (\( f \) is defined precisely if \( p \) is specified);

⁶ A strategy (based on “model averaging”) for the use of differing models, in order to provide inferences, is described by Burnham and Anderson ([22, p. 326] and [23, p. 448]).
\( \psi \) and \( \psi_1 \) are initial functions (possibly parameter-dependent) and for a choice of \( p \) the values \( y(t_j; p) \) with components \( y^i(t_j; p) \) will be expected to simulate data \( \{y_j\} \) (observed at times \( t_j \in [t_{\text{start}}, T] \)) with components \( \{y^i_j\} \) \((i = 1, 2, \ldots, M, j = 1, 2, \ldots, N)\). We can be asked to identify parameter values in \( \psi(t; p) \) and \( \psi_1(t; p) \) as well in the equations. The question of identifying the initial function \( \psi \), not broached here, is addressed in [12]. Throughout the paper,

1. \( L \) is the number of parameters,
2. \( M \) is the dimensionality of the state vector,

and \( \nu \) denotes the sample size (the number of scalar observations, usually \( \nu = NM \)). Models of the type (2.1) appeared in [10]; a form of NDDE known as Hale’s form is discussed in [14].

It may be asked where the mathematical models originate. There is a vast body of expertise in 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 bio-mathematics, we may cite [2,3,8,16,39,47,52] as examples of good practice. The book [2] addresses issues of compartmental modelling and [52] contains valuable biological information and reflections on modelling philosophy within theoretical regulatory biology.

2.1. Data

We presume that we have observations that we take as input data. It is possible that the data comes from a single experiment, but more likely that the data arises from several experiments or a series of observations. In this case, individual observations have to be summarized in some way. We assume 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.) of the errors in the observations, it is important to ensure that these assumptions are realistic.

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 \( \{\sigma^2_j\} \) about the mean values \( y^j_j \) at time \( t_j \). In this circumstance, one strategy is to choose the parameter \( p \) in the model defining \( y(t; p) \) in order to minimize

\[
\phi(p) := \sum_{i,j} \frac{1}{\{\sigma^2_j\}_i^2} |y^i(t_j; p) - y^i_j|^2,
\]

assuming normally distributed errors. This objective function represents the square of a norm of the distance of the vectors \( y(t_j; p) \) from the vectors of observations \( y_j \) associated with a certain set of observation times \( \{t_j\} \). There are echoes of the above procedure in what we propose as a methodology, but

\footnote{With little amendment we can consider the case where different components \( \{y^i_j\}_{i=1}^N \) are associated with \( i \)-dependent times \( \{t^i_j\}_{j=1}^N \).}
we are proceeding on the basis that we are provided with the data in the form of the mean values, not the individual observations.

2.2. Model identification

Model identification may be regarded as having two components: (i) selection, from a range of forms, of a parametrized mathematical model, and (ii) computation of the value of the parameter that yields 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, e.g., [41]). The perspectives that one may adopt are affected by the nature of the data. If the data is subject to observational errors of a precise statistical nature (independent, and having Gaussian distribution and zero mean)—concepts made precise in [36, Eqs. (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 [22, Chapter 2] and [23]. (Burnham and Anderson [23, p. 98] consider the theory and use of the information-theoretic approach to model building and data analysis to be simpler than that of the Bayesian approaches.)

2.3. Well-posedness

A fundamental question governing a mathematical problem is (cf. [50]) 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 in the term “continuous dependence” indicates that ill-posedness can only be discussed within a precise framework that includes definitions of the data and solution spaces. Hadamard appears to have considered 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 distinguish ill-posedness from 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, 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 type of discretization.

Remark 2.1. Various regularization techniques for ill-posed problems can be found in the literature [50]; 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 (see, e.g., (3.2) below) by minimizing a new objective function formed by the addition of a regularization term that depends on a regularization parameter. See, for example, [12]. For regularization and data assimilation in other

8 For the archetypal ill-posed linear equation of the form $K\phi = g$ where $K : X \to X$ is compact, the rate of decay to zero of the singular values of $K$ indicates the degree of ill-posedness.
contexts, see [38]. A regularization parameter is often introduced in data assimilation (cf. [12,42]); from one viewpoint, it may be regarded as a means for introducing information not implicit in the model. Links between regularization and maximum likelihood techniques can be found in the literature, but their application in the context of our present topic appears to deserve further attention; regularization is not covered in [23], for example.

We may distinguish the cases where

(i) we have 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) we have 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;
(iii) we have data that comprises mathematically defined entities (e.g., observations that are presented as an error-free densely-defined function—an idealized case);
(iv) we have 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) we have an aggregate of observations, on a discrete set of points, that may be subject to errors (errors that are, say, normally distributed—a case that is frequently assumed);

and so on. We must choose from amongst the scenarios itemized in (i)–(v) above. Thus, there may be no parameters that provide 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.

2.4. Theoretical identifiability

From the perspective adopted here, identifiability is concerned with whether, amongst a hierarchy of models, there is an identifiable model that optimizes some measure of best fit to noisy data (a measure that may include an index reflecting complexity or parsimony). However, 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” [6,18]. There, the assumption is that the form of the model has been correctly identified (the actual model is to be characterized by a parameter p), and the observational data \{y_j\} is assumed to be error-free. For us, our data is subject to noise, and the best model (taken from a hierarchy of models) is an outcome not an assumption. Nevertheless, we gain insight from the alternative perspective. Some examples will illustrate particular points.

Example 2.1. Consider \( y'(t) = p_0 y(t) \) for \( t \geq 0 \), with \( y(0) = p_1 \), which has the solution \( y(t) = p_1 \exp(p_0 t) \). If \( y(1) \) is an observed value, 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.
We give a further example based on DDEs; this shows, *inter alia*, that the design of the experiment (at least, the selection of the points \( \{ t_i \} \)) is a factor in determining identifiability. The following discussion shows that there may not be a unique parameter whatever the size of the sample set.

**Example 2.2.** Consider a simple model associated with the DDE: \( y'(t) = \rho_0 y(t) + \rho_1 y(t - 1) \ (t \geq 0) \), with \( y(t) = \exp(y t) \ (t \leq 0) \). If \( \rho_0 = 1, \ \rho_1 = 1 \) and \( \gamma = 0 \) we obtain the unique solution \( y(t) = 2 \exp(t) - 1 \) on \([0, 1]\). On the other hand, if \( \rho_0 = 0, \ \rho_1 = 2\gamma \) and \( \gamma = 1 \) we obtain the same unique solution \( y(t) = 2 \exp(t) - 1 \) on \([0, 1]\). Thus any exact data specified for an arbitrary set of arguments \( \{t_j\}_{j=1}^N \) in the interval \([0, 1]\) corresponds to two possible vectors of parameters, \([\rho_0, \rho_1, \gamma]^T \). The model is not uniquely identifiable for arbitrary sets of data in \([0, 1]\).

Now consider a simple model governed by the NDDE \( y'(t) = \rho_0 y(t) + \rho_1 y(t - 1) + \rho_2 y'(t - 1) \ (t \geq 0) \), with \( y(t) = \psi(t), \ y'(t) = \psi'(t), \) for \( t \leq 0 \). Suppose that we have values \( y_i = y(t_i) \) for \( i = 0, 1, \ldots, N \), where \( t_i \in (0, 1) \) and \( \psi(t) \equiv \alpha \). On \([0, 1]\), \( y'(t) = \rho_0 y(t) + \alpha \rho_1 \), and

\[
y(t) = \frac{\alpha \rho_1}{\rho_0} \left\{ \exp(\rho_0 t) \left\{ 1 + \frac{\rho_0}{\rho_1} \right\} - 1 \right\} \quad \text{on} \ [0, 1].
\]

Thus, \( \rho_2 \) is not determined and the parametrized model is not uniquely identifiable for arbitrary data defined only on \([0, 1]\); in this case, we can at best determine \( \rho_0 \) and the product \( \alpha \rho_1 \) rather than the separate values \( \alpha, \rho_0, \rho_1 \). If \( y(0) \) is known, then \( \alpha \) is determined, and if additional data is provided on the interval \([1, 2]\), \( \rho_2 \) could be found, etc.

In the latter example, we see "observational parameters" \( q \) with components \( q_0 = \rho_0 \) and \( q_1 = \alpha \rho_1 \) (or with \( q_0 = \rho_0 / \rho_1 \) and \( q_1 = \alpha \)) which define an infinite number of parameters \( p = [\rho_0, \rho_1, \alpha]^T \). The general question is whether one can proceed, and if so in a *unique* way, to determine observational parameters \( q \) from \( \{y_i\} \) and thence to obtain \( p \) from \( q \). The possibilities include: (i) the non-existence of suitable parameters; (ii) the existence of a unique parameter; (iii) the existence of infinitely many (non-unique) parameters; (iv) the existence of infinitely many parameters. Cobelli and his co-authors [6, 7, 25] 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 non-linear or even linear models, is however difficult, as it requires solving a system of non-linear equations which increases both in non-linearity and number of terms and unknowns with increasing model complexity; see [6, 7]. In our models, we endeavour to employ the model parameters themselves as the observational parameters.

### 2.5. Practical identifiability

The preceding discussion is based on a different perspective from that implicit in our practical approach. Our techniques for parameter estimation involve (*inter alia*) minimizing a function \( \Phi(p) \) of the parameters \( p \) that is constructed from the observed data. In general, for a given data set \( \{t_j, y_j\}_{j=1}^N \) and an arbitrary function \( f \) in (2.1), there is no reason to suppose that there exists a *unique* minimizer \( \hat{p} \) of \( \Phi(p) \). Indeed, it is easy to find examples of non-unique best fit models (seek solutions for two different parameters that agree at the points \( t_1, t_2, \ldots, t_N \)). The graphs (cf. [11, Part I]) of solutions of \( y'(t) = y(t)(a - y(t - 1)) \) (for \( t \geq 0 \); with \( y(t) = \psi(t) \) (for \( t \in [-1, 0] \)) where \( p = [a] \), and \( 1 \leq a < 1.6 \) (say), demonstrate that solutions for different parameters may pass through a common set of values. If the data correspond to the points of intersection, \( \hat{p} \) is not uniquely determined. Even if there exists a
unique $\hat{\mathbf{p}}$, the success of iterative methods for determining its value may depend upon a close starting approximation.

Practical identifiability of parameters, i.e., the uniqueness of the parameter vector $\hat{\mathbf{p}}$ minimizing the objective function $\Phi(\mathbf{p})$, is related to the “shape” of the objective function in terms of the parameters. As illustrated in Fig. 1 (based on experimental data), it is instructive to obtain graphical illustrations of the behaviour of $\Phi(\mathbf{p})$ in neighbourhoods of $\hat{\mathbf{p}}$. Valley-type behaviour of $\Phi(\mathbf{p})$ with a level valley floor (e.g., Fig. 1(a)) implies a high correlation between corresponding parameters: any combination of parameter components corresponding to the valley floor is equally effective at providing a good fit, as represented by a small objective function. This interpretation is supported by analysis of the covariance matrix for parameter estimates $\mathbf{S}(\mathbf{p})$ in (6.2) below.

The presence of errors in data sets affects the estimated parameters. It is desirable to specify the uncertainty (e.g., 95% confidence intervals) in the estimates by quantifying the effect of the data errors on the parameter estimates. Therefore, practical identifiability analysis involves examining, typically from the covariance matrix (see Section 6.2), the achievable accuracy of the parameter estimates obtained from the given noisy data.

3. Least-squares approach to the estimation of model parameters

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. We choose a suitable objective function $\Phi(\mathbf{p})$, which depends upon the given data $\{t_j; y^i_j\}_{j=1}^N$ (for $i = 1, \ldots, M$) and the values $\{y^i_j(t; \mathbf{p})\}_{j=1}^M$ of the solution $\mathbf{y}(t; \mathbf{p})$ of the parametrized model, e.g., (2.1). We seek the parameter $\hat{\mathbf{p}}$ for

Fig. 1. 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 non-linearity of the model.
for which the corresponding values \( \{ y^i(t_j; \hat{p}) \}_{j=1}^{1:M} \), provide a ‘best fit’ to the given data \( \{ y^j \}_{j=1}^{1:N} \) in the sense that

\[
\Phi(\hat{p}) = \min_p \Phi(p).
\]  

(3.1)

This problem is well-posed if \( \Phi(\cdot) \) has a unique minimizer. Now \( \Phi(p) \geq 0 \) and the size of \( \Phi(p) \) is in general a measure of the corresponding residual, that is, a measure of the amount by which the values \( y(1); p) \) differ from the values \( y_j \).

The gradient of a smooth objective function \( \Phi(p) \) is the vector \( \frac{\partial}{\partial p} \Phi(p) \) and the Hessian matrix is

\[
\mathbf{H}(p) := \{ \frac{\partial}{\partial p} \}^{T} \Phi(p) \in \mathbb{R}^{L \times L}, \ p \in \mathbb{R}^{L}, \text{ with } (i, j) \text{th element } \frac{\partial^2}{\partial p_i \partial p_j} \Phi(p), \text{ assuming that the derivatives exist.}
\]

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

**Remark 3.1.** *If the Hessian is positive semi-definite, the minimum referred to in Theorem 3.1 is not necessarily strict. Note that a positive-definite matrix* \( \mathbf{H} \) *has positive diagonal elements. We warn that with models based on (2.1) the objective function may not depend in a smooth manner on \( \tau \); this has implications if one seeks to apply standard computational techniques that are based on the assumption that the objective function is a smooth function of the parameter values. For remarks on differentiability with respect to parameters, see [13,30].

Usually, the solution \( y(t; p) \) 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 successive parameter estimates, and they require initial starting values. An obvious difficulty is that there is the possibility of the iteration converging to a local minimum, or not converging at all, rather than converging to the desired global minimum.

As a practical detail, a procedure for finding the minimizer \( \hat{p} \) of an objective function \( \Phi(p) \equiv \Phi(\{t_i\}_1^N; p) \) is to choose \( \ell \in \{1, 2, \ldots, L - 1\} \), fix parameter values \( \tilde{p}_{\ell+1}, \tilde{p}_{\ell+2}, \ldots, \tilde{p}_L \) and find \( \tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_\ell \in \mathbb{R}^{\ell} \) such that with \( \tilde{p}_\ell = [\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_\ell, \tilde{p}_{\ell+1}, \ldots, \tilde{p}_L]^T \),

\[
\Phi(\{t_i\}_1^N; \tilde{p}_\ell) = \min_{q \in \mathbb{R}^L} \Phi(\{t_i\}_1^N; q).
\]

We then increase \( \ell \) toward \( L \). This procedure corresponds to examining the \( \ell \)th model in a sequence of nested models and we may take \( \tilde{p}_\ell \) as an approximation to the vector parameter \( \hat{p} \) that minimizes \( \Phi(\{t_i\}_1^N; p) \). Alternatively, \( \hat{p} \) can often be approximated by the parameter that minimizes an objective function based upon a subset of the observations (at data points \( \{t_i\}_1^n \) with \( n < N \)). Let us write this objective function \( \Phi_n(\{t_i\}_1^N; p) \) (for \( p \in \mathbb{R}^{L} \)) and denote by \( \hat{p}_n \) the vector such that \( \Phi_n(\{t_i\}_1^N; \hat{p}_n) \leq \Phi_n(\{t_i\}_1^N; \hat{p}) \) for all \( p \in \mathbb{R}^{L} \). With sequentially extended objective functions \( \{ \Phi_n(\{t_i\}_1^n; p) \}_{n=1}^{N} \) we have \( \hat{p}_N \approx \hat{p}_N =: \hat{p} \) for \( n \) close to \( N \). Also, \( \hat{p}_{n-1} \) provides a starting approximation when determining \( \hat{p}_n \) (\( n = 1, 2, \ldots, N \)).
Remark 3.2. 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. A methodology based on this strategy is inadequate as it excludes significant modelling principles.

With regard to the number of parameters in the model, Burnham and Anderson [23] comment “In general, bias decreases and variance increases as the dimension of the model increases. The fit of any model can be improved by increasing the number of parameters, however, a trade-off with the increasing variance must be considered in selecting a model for inference.”

3.2. Least-squares and related objective functions

In our discussion of (3.1), we first focus on ordinary least squares (OLS) and weighted least squares (WLS) fitting. This corresponds to objective functions $\Phi_{\text{OLS}}(\{t_i\}_1^N; p) \equiv \Phi_{\text{WLS}}(\{t_i\}_1^N; p) \equiv \Phi_{\text{WLogLS}}(\{t_i\}_1^N; p)$, respectively, where

\[
\Phi_{\text{OLS}}(p) := \sum_{j=1}^N \sum_{i=1}^M [y'(t_j; p) - y_j^j]^2 := \sum_{j=1}^N \|y(t_j; p) - y_j\|^2; \quad (3.2a)
\]

\[
\Phi_{\text{WLS}}(p) = \sum_{j=1}^N \sum_{i=1}^M w_i^j [y'(t_j; p) - y_j^j]^2. \quad (3.2b)
\]

(Generally, $\Phi(p)$ is associated with a (semi-)norm of the residual; consider $\sqrt{\Phi_{\text{WLS}}(p)}$, for example.) We also consider weighted log-least-squares (WLogLS) fitting. This corresponds to an objective function $\Phi_{\text{WLogLS}}(\{t_i\}_1^N; p) \equiv \Phi_{\text{WLogLS}}(p)$ where

\[
\Phi_{\text{WLogLS}}(p) = \sum_{j=1}^N \sum_{i=1}^M w_i^j [\ell n(y'(t_j; p)) - \ell n(y_j^j)]^2, \quad (3.2c)
\]

where we denote the natural logarithm by $\ell n(\cdot)$. To use (3.2c), it will be assumed that $y_j^j > 0$ and $y'(t_j; p) > 0$. The objective function (3.2c) may be interpreted as a realization of (3.2b) in which the data are regarded as the values $\ell n(y_j^j)$ and we consider a model (not necessarily of standard type) that has a solution $\ell n(y(t_j; p))$; thus, in the case of (3.2c), any assumptions about the data must be valid for the values $\ell n(y_j^j)$. Important observations are given by Burnham and Anderson [23, p. 81], which amount to the need for consistency in one’s approach. Various objective functions correspond [28] 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 (3.2a) or (3.2b)), 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 (3.2c).

Remark 3.3. If $\tau$ is to be estimated in a model of the form (2.1), then $\tau$ is a component of $p$ that is constrained to be non-negative, and (3.1) is clearly a constrained minimization problem. However, since one can replace $\tau$ by $p_{\mu}^2$ for an unconstrained parameter $p_{\mu}$, the problem (3.1) is in principle an unconstrained minimization. On the other hand, it is in practice wise to introduce well-founded constraints on
admissible parameters. In cell growth, a parametrized model is unrealistic if the corresponding predicted populations are found to be negative.

The notation $\Phi_*(p)$ will be taken to denote a generic objective function that can be any one of (3.2a)–(3.2c). The objective function $\Phi_{WLS}(p)$ can be expressed as

$$\sum_{j=1}^{N} \left( [y(t_j; p) - y_j]^T W_j [y(t_j; p) - y_j] \right) \equiv \sum_{j=1}^{N} \| W_j^{1/2} [y(t_j; p) - y_j] \|^2,$$

on defining a set $\mathcal{W} = \{ W_j \}$ of matrices that we usually suppose are diagonal, i.e.,

$$W_j = \text{diag}[w_j^1, w_j^2, \ldots, w_j^M],$$

where $w_j^i > 0$. The choice of positive weights $w_j^i$ can be based on knowledge of the relative accuracy of the individual components of $y_j$. (One can permit $w_j^i = 0$ if the corresponding data value, the $i$th component $y_j^i$ of $y_j$, is missing.)

In principle, there is nothing to prevent the use of a set $\mathcal{W}$ in which each matrix $W_j$ is positive-definite, to define $\sum_{j=1}^{N} \| W_j^{1/2} [y(t_j; p) - y_j] \|^2$ as an objective function generalizing (3.2b). We shall concentrate on positive-definite diagonal matrices (3.3b). 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 $\sigma_j^2$ is the variance:

$$\sigma_j^2 := \mathbb{E}(\{ \mathcal{E}(y_j^i) - y_j^i \}^2) \quad \text{(assumed independent of } i),$$

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

$$W_j = \sigma_j^{-2} I;$$

in the case $\sigma_j = \sigma$ for all $j$, this becomes $W_j = \sigma^{-2} I$ where $I$ is the identity. As a further example of (3.3b), we may propose $w_j^i = (\sigma y_j^i)^{-2}$, where $\sigma$ is termed the “coefficient of variation” (CV)—see Landaw and DiStefano [36]—to arrive at

$$W_j = \sigma^{-2} \text{diag}^{-2}[y_j^i].$$

Remark 3.4. If we have two collections of (positive-definite) weighting matrices $\mathcal{W}' = \{ W'_j \}$, and $\mathcal{W}'' = \{ W''_j \}$ where $W'_j = s^2 W''_j$ for some non-zero $s$; then the minimizer $\hat{p}$ of $\Phi_{WLS}(p)$ is clearly the same as the minimizer of $\Phi_{W_{WLS}}(p)$ and from this viewpoint the factor $s^2$ is irrelevant. However, $s$ is relevant in terms of the size of the objective functions at the minimum, and, for an appropriate choice of weights $\mathcal{W}$, the minimum value $\Phi_{WLS}(\hat{p})$ has a rôle in the evaluation of certain indicators of the acceptability of the model.

For log-least-squares, we require some further notation. Given $z = [z_1, z_2, \ldots, z_m]^T \in \mathbb{R}^m$ we employ the notation $\ln(z)$, when $z$ has positive components $(z > 0)$, to denote $\ln(z) = [\ln(z_1), \ln(z_2), \ldots, \ln(z_m)]^T$. We can then take for $\Phi_{WLogLS}(p)$ the expression

\footnote{The notation $\text{diag}'[z^T]$ signifies the diagonal matrix diag$\{z_1^r, z_2^r, \ldots, z_m^r\}$, for $r \in \mathbb{R}$—if $z > 0$ and for $r \in \mathbb{R}^+$ for all $z$; it is the identity matrix if $r = 0$.}
Under Assumption 4-1, the probability density functions are given by (the latter quantities being assumed independent of)

\[ \text{Assumption 4-2} \]

for which the likelihood is the highest are the “maximum likelihood estimates”. The data is regarded as fixed and the parameters as variable; the parameters observed data are assumed to be expressible using an appropriate probability distribution (e.g., a normal, or log-normal, distribution). The data is regarded as fixed and the parameters as variable; the parameters for which the likelihood is the highest are the “maximum likelihood estimates”.

An appropriately chosen objective function \( \Phi^*(p) \) based on properly chosen weights can provide an approximation to the maximum likelihood estimator under certain assumptions, in particular (Assumption 4-1) that the errors in observations at successive times are independent. If (Assumption 4-2) the errors in the observed data are assumed to have a Gaussian distribution (e.g., a normal, or log-normal, distribution). The data is regarded as fixed and the parameters as variable; the parameters for which the likelihood is the highest are the “maximum likelihood estimates”.

We start with an informal sketch of the maximum likelihood formulation. This allows one to find the value of a parameter that maximizes the probability (likelihood) of obtaining exactly the observed data (see [5,17,40], etc.) using the computed parameter. The distributional properties of the errors in the observed data are assumed to be expressible using an appropriate probability distribution (e.g., a normal, or log-normal, distribution). The data is regarded as fixed and the parameters as variable; the parameters for which the likelihood is the highest are the “maximum likelihood estimates”.

An appropriately chosen objective function \( \Phi^*(p) \) based on properly chosen weights can provide an approximation to the maximum likelihood estimator under certain assumptions, in particular (Assumption 4-1) that the errors in observations at successive times are independent. If (Assumption 4-2) the errors in the observed data are assumed to have a Gaussian distribution about the vectors \( \{z_j\}_{j=1}^N \), that is, \( y_j \sim N(z_j, \Sigma_j) \), where \( \Sigma_j \) is the \( j \)th covariance matrix

\[
\Sigma_j := \begin{bmatrix}
  (\sigma_{11}^{(j)})^2 & \rho_{12}^{(j)} \sigma_{11}^{(j)} \sigma_{12}^{(j)} & \cdots & \rho_{1M}^{(j)} \sigma_{11}^{(j)} \sigma_{1M}^{(j)} \\
  \rho_{21}^{(j)} \sigma_{21}^{(j)} \sigma_{22}^{(j)} & (\sigma_{22}^{(j)})^2 & \cdots & \rho_{2M}^{(j)} \sigma_{21}^{(j)} \sigma_{2M}^{(j)} \\
  \vdots & \vdots & \ddots & \vdots \\
  \rho_{M1}^{(j)} \sigma_{M1}^{(j)} & \rho_{M2}^{(j)} \sigma_{M2}^{(j)} & \cdots & (\sigma_{MM}^{(j)})^2
\end{bmatrix}, \quad \text{with } \rho_{lk}^{(j)} = \frac{\text{Cov}(y_{l}^{(j)}, y_{k}^{(j)})}{\sqrt{\text{Cov}(y_{l}^{(j)}, y_{l}^{(j)}) \text{Cov}(y_{k}^{(j)}, y_{k}^{(j)})}}
\]

(\( \rho_{lk}^{(j)} \) denoting the correlation coefficient of the \( l \)th and \( k \)th observed component of \( y_j \)) then the component probability density functions are given by

\[
\mathcal{H}(y_j; p) = \frac{1}{\sqrt{(2\pi)^{M}}} \exp\left\{ -\frac{1}{2} [z_j - y_j]^{T} \Sigma_j^{-1} [z_j - y_j] \right\}^{N}_{j=1}.
\]

Under Assumption 4-1 and Assumption 4-2, the likelihood function is given by

\[
\mathcal{L}(p) = \prod_{j=1}^{N} \mathcal{H}(y_j; p).
\]
where \( \mathcal{H}(y_j; p) \) is the probability density function (4.2). Thus, \( \mathcal{L}(p) \) is a function of the matrices \( \Sigma_j \), but if the observation errors at successive times were not independent, then the \( N \) matrices (4.2) would be inadequate to characterize the stochastic process. It would then be necessary to revise (4.3), but the changes are straightforward if \( M = 1 \) and in this case the correlations (cf. [21]) enter into the amendments of our subsequent formulae.

Remark 4.1. The probability density function (4.2) and the resulting likelihood function (4.3) correspond to a model assuming a normal distribution of the errors in the observations. In the case of a log-normal distribution, the probability density function and, therefore, the likelihood function, take a slightly different form (for details, see [23, pp. 82 & 318]).

If (Assumption 4-3) the errors in the components of \( y_j \) are assumed to be independent then \( \rho^{(i)}_{\ell k} = 0 \) (for \( \ell \neq k \)). In this case,

\[
\Sigma_j := \text{diag}\left(\left(\sigma^{(i)}_{\ell j}\right)^2, \left(\sigma^{(j)}_{\ell j}\right)^2, \ldots, \left(\sigma^{(M)}_{\ell j}\right)^2\right). \tag{4.4}
\]

For theoretical and practical convenience it is often preferable to work in terms of the logarithm of the likelihood function. If we now suppose (Assumption 4-4) that \( z_j \) is the value \( y(t_j; p) \) then, given (4.4),

\[
-2\elln \mathcal{L}(p) = NM\elln(2\pi) + 2\elln\left(\prod_{i,j} \sigma^{(i)}_{\ell j}\right) + \sum_{j} \|\Sigma_j^{-1/2}[y(t_j; p) - y_j]\|^2. \tag{4.5}
\]

If the weights in \( \Phi_{\text{WLS}}(p) \) are chosen to match the variance of the error, so that \( W_j = \left\{\text{diag}[\sigma^{(1)}_{\ell j}, \sigma^{(2)}_{\ell j}, \ldots, \sigma^{(M)}_{\ell j}]\right\}^{-2} \) in (3.2b), then the last term in (4.5) becomes \( \Phi_{\text{WLS}}(p) \). In particular, if \( W_j = \Sigma_j^{-1} \), and \( \sigma_i = \sigma \) for all \( i \), then (Assumption 4-5)

\[
W_j = \sigma^{-2}\Omega_j, \quad \text{where} \quad \Omega_j = \left\{\text{diag}\left[\omega^{(1)}_{\ell j}, \omega^{(2)}_{\ell j}, \ldots, \omega^{(M)}_{\ell j}\right]\right\}^{-2} \tag{4.6a}
\]

(compare (3.4b)) then \( \Phi_{\text{WLS}}(p, \sigma) \equiv \Phi_{\text{WLS}}(p, \sigma) \) and

\[
\Phi_{\text{WLS}}(p, \sigma) = \sigma^{-2}\Phi_{\text{GLS}}(p). \tag{4.6b}
\]

This is to say \( \Phi_{\text{WLS}}(p, \sigma) \equiv \sigma^{-2}\sum_j \|\text{diag}^{-1}[\omega^{(1)}_{\ell j}, \omega^{(2)}_{\ell j}, \ldots, \omega^{(M)}_{\ell j}][y(t_j; p) - y_j]\|^2 \). Then (4.5) yields for

\[
-2\elln \mathcal{L}(p) = -2\elln \mathcal{L}(p; \sigma) \text{ the expression}
\]

\[
-2\elln \mathcal{L}(p; \sigma) = NM\elln(2\pi) + NM\elln(\sigma^2) + 2\sum_{i,j} \elln(\omega^{(i)}_{\ell j}) + \sigma^{-2}\Phi_{\text{GLS}}(p), \tag{4.6c}
\]

where \( \Omega = \{\Omega_j\} \). The aim in the maximum likelihood strategy is to maximize (4.3), or equivalently (according to (4.6c)) to minimize (4.6b) by an appropriate choice of \( p \) and \( \sigma \). Actually, the optimal value \( \hat{\sigma} \) of \( \sigma \) follows from the optimal value \( \hat{p} \) of \( p \) and one may proceed by minimizing \( \Phi_{\text{GLS}}(p) \). Indeed, using the necessary optimality condition, \( \frac{\partial (\elln \mathcal{L}(p; \sigma))}{\partial \sigma} = 0 \), to provide the maximum likelihood estimate, we obtain the estimate

\[
\hat{\sigma}^2 = \frac{1}{NM} \sum_j \|\text{diag}^{-1}[\omega^{(1)}_{\ell j}, \omega^{(2)}_{\ell j}, \ldots, \omega^{(M)}_{\ell j}][y(t_j; \hat{p}) - y_j]\|^2 = \frac{1}{NM} \Phi_{\text{GLS}}(\hat{p}). \tag{4.7}
\]
Remark 4.2. Suppose Assumption 4.5 is valid, and $W_j = \{\text{diag}^{-2}[\sigma_1, \sigma_2, \ldots, \sigma_M]\}$ (so $W_j$ is independent of $j$). Then in place of (4.7) we estimate

$$
\hat{\sigma}_i^2 = \frac{1}{N} \sum_j |y^j(t_j; \hat{p}) - y^j_j|^2,
$$

(4.8)

where $\hat{\sigma}_i^2$ is the maximum likelihood estimator of the variance in the $i$th component of the $j$th observed vector (constant for all $j$). The total number of parameters estimated is now $L + M$.

The quantity (4.5) (which is a multiple of the natural logarithm of the maximized likelihood) now becomes

$$
-2\ln L(\hat{p}; \sigma) = \left\{ NM\ln(2\pi) + NM + 2\sum_{i,j} \ln(\omega^{ij}_j) \right\} 
+ NM\ln(\Phi_{\text{OLS}}(\hat{p})) - NM\ln(NM).
$$

(4.9)

Remark 4.3. The standard least-squares estimator (see [22, p. 17]) of $\sigma^2$ differs from (4.7) 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. [31] 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 $\sigma^2$, $\Sigma_j = \sigma^2 I$.

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 $W_j = \Sigma_j^{-1}$, 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 these weights.

4.1. Confidence intervals

Approximate (e.g., 95%) confidence regions for every component parameter $p_\ell$ of the optimum vector $\hat{p} \equiv [p_1, p_2, \ldots, p_\ell, \ldots, p_L]^T$ can be obtained using the profile likelihood method.\(^{10}\) This proceeds as follows: Suppose that we concentrate upon the $\ell$th component $\hat{p}_\ell$ of the optimal parameter vector $\hat{p} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{\ell-1}, \hat{p}_\ell, \hat{p}_{\ell+1}, \ldots, \hat{p}_L]^T$. We search for the interval $[p_\ell^\text{min}, p_\ell^\text{max}]$ of maximal width and containing $\hat{p}_\ell$ such that, with

$$
\mathcal{L}_\ell(\hat{p}) = \max_{p\in S_\ell(p)} \mathcal{L}(p), \quad \text{where } S_\ell(p) := \{[p_1, p_2, \ldots, p_{\ell-1}, p, p_{\ell+1}, \ldots, p_L]^T | p \text{ fixed} \},
$$

we have

$$
|\ln(\mathcal{L}_\ell(\hat{p})) - \ln(\mathcal{L}(\hat{p}))| \leq \frac{1}{2} \chi^2_{1.0.95}, \quad \text{whenever } p_\ell \in [p_\ell^\text{min}, p_\ell^\text{max}],
$$

(4.10)

\(^{10}\) See [9] where two other methods are mentioned: (a) the variance-covariance matrix based technique (which requires that the objective function be ‘small’), (b) the bootstrap methods (whose validity has been challenged from a probabilistic aspect).
where $\chi^2_{1,0.95}$ stands for the 0.95th quantile of the $\chi^2$-distribution for 1 degree of freedom. Using the relationship (4.9) between the MLE and the least squares estimation we obtain

$$NM \left| \ell n(\Phi/\Omega_1^{LS}(\hat{p})) - \ell n(\Phi/\Omega_1^{LS}(\hat{p})) \right| \leq \chi^2_{1,0.95}, \quad \text{whenever } p_\ell \in [p_\ell^{\min}, p_\ell^{\max}].$$

For one particular data set the total number of scalar observations ($NM$ in the above formula) is 15, and the tabulated value of $\chi^2_{1,0.95}$ is 3.841. Hence, the final expression for computing the 95% confidence interval reads

$$\left| \ell n(\Phi/\Omega_1^{LS}(\hat{p})) - \ell n(\Phi/\Omega_1^{LS}(\dot{p})) \right| \leq \frac{3.841}{15}.$$

**Remark 4.4.** A comparison of variance-covariance, profile-likelihood and bootstrap methodologies (each of which is justified under particular assumptions) is planned by the authors for a later paper.

5. Model evaluation based on an information-theoretic approach

When one has confidence in the form of the model, the goodness of fit associated with parameter estimates $\hat{p}$ can be characterized by the size of an objective function $\Phi_*(\hat{p})$. This is the data-fitting approach, and here $\hat{p}$ may be an approximation to $\hat{p}$ such that $\Phi_*(\hat{p}) = \min_p \Phi_*(p)$. Thus, one criterion by which to judge a model may be the size of $\Phi_*(\hat{p})$ (see [20]). 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 [40], whereas least squares estimation in terms of an objective function $\Phi_*(p)$ 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. (Other approaches such as the Bayesian approach can be of value, but we shall not discuss them here.) A further 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 is difficult or impossible to formalize.

There are statistical criteria (such as the Akaeke, Schwarz, and Takeuchi information criteria and generalizations related to informational complexity of models), which depend not only upon estimates of the maximum likelihood estimator (MLE) [1,5,48] but incorporate the number of parameters and the number of observations, for a quantitative evaluation of different models. The criterion of Akaeke is related both to the maximum likelihood estimate (MLE) [48] and to the Kullback–Leibler (K–L) notion [34] of “distance” between two models (an information-theoretic concept). Burnham and Anderson [22,23] review both the concept of K–L information as a natural basis for model selection (“a dominant paradigm in information and coding theory”), and maximum likelihood (“the dominant paradigm in statistics”). They describe the Akaeke criterion, which Akaeke related to the K–L theory in 1973, as “a new paradigm in statistical science”. The Akaeke Information Criterion (AIC) [22, §2.2], [23, §2.2] and the corrected Akaeke Information Criterion (cAIC) [22, §2.4], [23, §2.4] (as well as other criteria such as that of Takeuchi [22, p. 67], [23, p. 65] and that of Schwarz [22, p. 68] not emphasized here) produce “indicators” that have been used as criteria for model selection [36,37,41,51,53]. 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 Akaeke criterion see [36,37].
We suppose $L(\hat{p})$ is the maximized likelihood function, and recall that $L$ is the number of free parameters, $M$ is the dimensionality of the state vector, and $v$ is the sample size (the number of scalar observations); in the present discussion $v = NM$. For the Akaike and the corrected Akaike criteria, the indicators expressed in terms of the MLE are, respectively, the size of the measures $\mu_{AIC}$ and $\mu_{cAIC}$ given by

$$\mu_{AIC} = -2\ln L(\hat{p}) + 2(L + 1),$$

$$\mu_{cAIC} = -2\ln L(\hat{p}) + 2(L + 1) + \frac{2(L + 1)(L + 2)}{v - L - 2}, \quad \text{with } v = NM,$$

respectively; see [22,23]. The number of parameters being estimated is $L + 1$, comprising $p_1, p_2, \ldots, p_L$ and $\sigma$ (we assume that a single value $\sigma$, which we estimate, characterizes all the variances). If a computed parameter $\hat{p}_i$ vanishes, this does not reduce $L$. The advice of Burnham and Anderson in [22, p. 322], [23, p. 66] is that (5.1a) is satisfactory if $v < 40(L + 1)$, otherwise (5.1b) is preferred by these authors. As $v \rightarrow \infty$, $\mu_{cAIC} \rightarrow \mu_{AIC}$.

**Remark 5.1.** In some scenarios, the data is divided 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 $v_i$) the correction term in (5.1b) consists of a sum over $i$ of terms in each of which $v_i$ replaces $v$; see [22, p. 255], [23, p. 379].

**Theorem 5.1.** In the case of normally distributed independent errors in the observations, Eqs. (5.1a) and (5.1b) can be expressed in terms of weighted least squares estimation (using appropriate weights that reflect the variance of the errors in the observations) as:

$$\mu_{AIC} = \left\{ NM\ln(2\pi) + NM + 2\sum_{i,j} \ln(\omega_{ij}) - NM\ln(NM) \right\} + NM\ln(\Phi_{QLS}(\hat{p})) + 2(L + 1),$$

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

(5.2a)  

(5.2b)

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

$$\tilde{\mu}_{AIC} = NM\ln(\Phi_{QLS}(\hat{p})) + 2(L + 1),$$

$$\tilde{\mu}_{cAIC} = \tilde{\mu}_{AIC} + \frac{2(L + 1)(L + 2)}{NM - L - 2}.$$  

(5.3a)  

(5.3b)

The values $\tilde{\mu}_{AIC}$ and $\tilde{\mu}_{cAIC}$ that are assigned to the AIC and cAIC 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) $\Phi_{QLS}(\hat{p})$. The best model is often taken to be that which yields the lowest value of $\mu_{AIC}$ or $\mu_{cAIC}$. This happens to take account of parsimony, to some extent. However, the terms often do not provide—at the least in the context of the models we discuss—a true measure of computational complexity of the model. What is complexity?
The formal measurements of complexity normally found in the literature are governed by the number of parameters—which does not reflect the abundance of qualitative diversity (e.g., those arising from non-linearity).

In modelling with DDEs, NDDE (or PDES), the initial or boundary functions defining a solution may depend upon some or all of the parameters, whereas for a model governed by an ODE the initial condition relates to an initial value—which often provides less rich dynamics. A DDE with one parameter is effectively infinite-dimensional whereas an ODE with one parameter is 1-dimensional. Since a DDE or NDDE with a fixed lag can (by the method of steps) be converted to systems of ODEs 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.

6. Sensitivity of the modelling process

Sensitivity analysis in modelling has been discussed elsewhere. It is appropriate here to cite Rabitz [45] who, in a different context, enumerates general applications of sensitivity analysis (without which a modelling process is “seriously deficient”), under the headings (i) elementary (first-order) sensitivities and higher-order sensitivities, (ii) parameter-observation interdependence (sensitivity to data), (iii) parameter interdependence, (iv) interdependence of different observations, (v) functional sensitivity analysis, (vi) memory effects and sensitivity analysis, (vii) position and time as dependent variables, (viii) sensitivity of objective functions, (ix) sensitivity to missing model components, (x) structural sensitivity analysis, (xi) mapping out parameter space, (xii) statistical error analysis.

If one considers practically significant perturbations, one may compute the differences (for example, \( y(t, p + \delta p) - y(t, p) \)). A study of the effect of ‘infinitesimal’ perturbations leads to an examination of derivatives (where they exist) that define the sensitivity coefficients, such as \( \frac{\delta}{\delta p} y(t, p) \). We shall indicate a use for such derivatives, below. The scaling of the parameter components \( p_i \) as they are represented in the model ((2.1), say) is a practical feature that may be important.\(^{11}\) Sensitivity, 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 indicate 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.

6.1. Non-linearity and indications of bias

The importance of the question of linearity versus non-linearity lies with the fact that the justification of various computations relies in places upon linear regression theory. For more details about non-linearity effects in parameter estimations, we may refer to [24].

\(^{11}\) The study of the relative sensitivity coefficients may prove helpful when scaling is an issue.
Bias of an estimator $\hat{p}$ is the difference between the expected value of the estimator and the true value, namely $b(\hat{p}) := E(\hat{p}) - p$; thus the estimator is unbiased if $b = 0$. Non-linear 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 [46, p.13] non-linear regression models have these properties only asymptotically (when the sample size becomes very large). The bias in the parameters depends on the nature of the non-linearity of the structural model [36,49] and can be examined numerically.

Following Ratkowsky [46, §2.6], given a weighted least-squares parameter estimate $\hat{p}$, one may proceed as follows:

- Perturb the solution values $y(t_j, \hat{p})$ of the model (corresponding to the best-fit parameters $\hat{p} \in \mathbb{R}_L$) by adding independent errors $\epsilon_i^j \sim N(0, \sigma^2)$ (see [17]), $\sigma^2 = \frac{\Phi(\hat{p})}{N-M}$, wherein the total number of observations is $N$ and $L$ denotes the total number of variable parameters;
- Find new best-fit parameter $\tilde{p}$ to the perturbed data $\{y^i(t_j, \hat{p}) + \epsilon_i^j\}$ (where $1 \leq j \leq N$ and $1 \leq i \leq M$);
- Repeat this process sufficient times (Ratkowsky [46] advises 1000 times) to generate a statistically significant estimate of the mean value $E(\tilde{p})$ of $\tilde{p}$.

If the relative bias is below a certain, small, threshold $c (\|\hat{p} - E(\tilde{p})\| < c\|\hat{p}\|)$, then both the bias of the weighted least-squares parameter estimate and the effect of non-linearity are deemed not to be significant [11, Part III].

### 6.2. Improvement of the data sampling strategy

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.

As Landaw and DiStefano observed [36], ad hoc sampling, particularly using equally spaced sampling points (which is conventional), can be inefficient in terms of the accuracy of the parameter estimates. The same precision can sometimes be obtained with fewer points, using D-optimal design [26]. With D-optimal design, we minimize $\det \mathbf{Z}(\hat{p})$ over the possible choices of $\{t_j\}_{j=1}^N$, where, for given $p$,

$$
\mathbf{Z}(p) \equiv \mathbf{Z}\left(\{t_j\}_{j=1}^N; p\right) \in \mathbb{R}^{L \times L} \quad \text{is the matrix with (r,s)th element } z_{rs} = \text{Cov}(p_r, p_s). \quad (6.1)
$$

The matrix (6.1) provides a measure of the dispersion of the best-fit estimate about its mean: If the estimator is unbiased (if $E(\hat{p}) - \hat{p}$ is small) then $\mathbf{Z}(\hat{p})$ characterizes the spread of the multivariate distribution around $\hat{p}$. For $\mathbf{Z}(\hat{p})$ we have

$$
\mathbf{Z}(\hat{p}) = 2 \frac{\Phi(\hat{p})}{(NM - L)} \times [H(\hat{p})]^{-1}, \quad (6.2)
$$

where $H(\hat{p})$ is the Hessian matrix. The Hessian matrix can be approximated in terms of the “information matrix” of the objective function $\Phi(\hat{p})$. Indeed, using the sensitivity coefficients $S(t; p) \equiv \left\{ \frac{\partial}{\partial p} y(t; p) \right\}$

we have
\[ H(\hat{p}) \approx 2 \sum_{j=1}^{N} S^T(t_j, \hat{p}) W_j S(t_j, \hat{p}) =: \tilde{H}(\hat{p}). \]
(The latter matrix, \( \tilde{H}(\hat{p}) \), is the information matrix.) Otherwise, we can calculate an exact value of \( \det \mathbf{S}(\hat{p}) \) from the first and second order sensitivity coefficients.

7. Computational aspects

7.1. Numerical tools for differential equations

Evaluation of the objective function \( \Phi(\hat{p}) \) requires us to solve a DDE or a NDDE (or, indeed, a PDE) with given parameters which we need to revise in an iterative manner in order to determine \( \hat{p} \). An analytical approach, to computing \( \Phi(\hat{p}) \), 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, e.g.,

(i) the NAg library http://www.nag.co.uk,
(ii) HSL: http://www.cse.clrc.ac.uk/nag/hsl—formerly the “Harwell Library”,
(iii) IMSL http://www.imsl.com, NETLIB (http://www.netlib.org) libraries, and

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 [43]. (For NDDEs the question of the continuity, existence and uniqueness of solutions is more complicated [14] than in the case of DDEs. A “proof-of-concept” code for discontinuous solutions of NDDEs in Hale’s form appears in [14].) Links to various codes can be found at

http://www.maths.man.ac.uk/~chris/software;
http://www.unige.ch/math/folks/hairer/software.html;
http://www.stats.gla.ac.uk/simon/simon/dde.html;

Clearly, different codes have different strengths or weaknesses. We refer to [43] for remarks on the design and analysis of numerical methods for DDEs and NDDEs and we draw particular attention to the codes [44]. Our numerical work based upon DDEs and NDDEs has relied on one-step continuous RK methods; Archi [44] is a versatile and robust code for a range of problems that are “non-stiff”. Various attempts have been made to distinguish between “stiff” and “non-stiff” DDEs; a distinction which even for ODEs has been regarded as not robust. 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.

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.
7.2. Minimization of objective functions

There is an extensive literature on optimization. Given the objective function $\Phi(p)$, we seek the optimum parameter $\hat{p}$ such that $\Phi(\hat{p}) \leq \Phi(p)$ for all physically meaningful values of $p$ and $\hat{p}$. The technique for finding the best-fit parameter values for a given mathematical model and objective function, given a set of data, consists of the following:

1. Providing an initial guess $p_0$ for the parameter estimates. This may arise from contour plotting or from estimates based on a reduced set of data or fewer parameters.
2. Solving the model equations using the current parameter values, to compute $\Phi(p)$.
3. Adjusting the parameter values (by the minimization routine, for example, EO4USF from the NAG library, now succeeded by EO4USF, or LMDIF from NETLIB or FMIN from MATLAB) so as to reduce the value of the objective function; see [44].

**Remark 7.1.** EO4USF, in Mark 20 of the NAG Library, is designed to minimize a smooth sum of squares function subject to constraints (simple bounds, or linear or smooth non-linear constraints) using a sequential quadratic programming (SQP) method. LMDIF is an unconstrained minimization routine based on the Levenberg–Marquardt algorithm. A version with a reduced parameter set, LMDIF1, is also available; in our experience, the default value of $\text{epsfcn}$ set internally in LMDIF1 should be amended when the model equations are solved numerically.

In the search for the optimum parameter, some minimization routines seek the evaluation of $\Phi(p)$ for “unrealistic” parameter values (for example, with $\tau < 0$ in (2.1)), which corresponds to an advanced rather than a delayed equation, or with parameters that give an unnecessarily “stiff” problem. We have encountered unsatisfactory situations of this type in one routine for constrained optimization, since (rather unexpectedly, in our view) the routine permitted violation of constraints in intermediate calculations.

Minimization routines may be classified as “derivative-free” or “gradient-dependent”. Contour-plotting can be made the basis of a derivative-free search routine, and has its value as an adjunct to other procedures. 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 additional computational expense.) 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 [13].

Our attempts to develop satisfactory models have led us to issues in the numerics of NDDEs. Solutions of DDEs and NDDEs can exhibit discontinuities or derivative discontinuities in [14,27], etc., that must be handled correctly in order to develop efficient and robust software [43]. (Analogous difficulties can arise in PDEs.) For the need for methods that deal with discontinuous solutions of NDDEs in Hale’s form, see [14]. Sometimes (cf. [13]), $\Phi_*(p)$ may not possess the smoothness properties required by many techniques for minimizing $\Phi_*(p)$. We try to check the best-fit parameters values by examining (with difficulty, in many dimensions) contour plots of the objective function.
8. Conclusions

The models with time-lag are at the complex end of a hierarchy of candidate models (we have not considered PDEs), from which we seek an “acceptable” parametrized model. The hierarchical feature is an essential part of our discussion. The complexity introduced by introducing a time-lag (which is strongly suggested by the science in the cases that we have studied) enforces a computational approach. In selecting a parametrized model to give quantitative and qualitative agreement with processes giving rise to observed data, there is a balance to be achieved between agreement with the data (under-fitting and over-fitting) and the principle of parsimony. What we propose relies 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 information theory, maximum likelihood, and a weighted least-squares fit, 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 the model and on the need to include all parameters.

The current paper is one of a sequence; it provides the essential methodology required in our related papers. Illustrative examples are provided in [10,12,14]; in [9], we endeavour, in a case study, to overcome real practical difficulties often ignored in “sanitized” expositions; further material appears in [11, Parts II, III]. While the present paper contains a number of original observations, we hope that it has a value greater than the sum of its parts (a number of which are known, but not to the whole research community); it attempts to integrate different perspectives. We advocate an interdisciplinary approach, and further attention to areas (in the theory and in the practice) where there are gaps in understanding. The computational approach requires the participation of numerical analysts, and we emphasize the need for their involvement, with other specialists, in what continues to be an interesting and challenging topic.

References