Infectious Disease Modelling 5 (2020) 111–128

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

A primer on model selection using the Akaike Information Criterion

Stéphanie Portet

Department of Mathematics, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada

ARTICLE INFO

Article history:
Received 11 December 2019
Accepted 27 December 2019
Available online 7 January 2020
Handling editor: Dr. J Wu

Keywords:
Collection of models
Model calibration
Model selection
Akaike information criterion

ABSTRACT

A powerful investigative tool in biology is to consider not a single mathematical model but a collection of models designed to explore different working hypotheses and select the best model in that collection. In these lecture notes, the usual workflow of the use of mathematical models to investigate a biological problem is described and the use of a collection of model is motivated. Models depend on parameters that must be estimated using observations; and when a collection of models is considered, the best model has then to be identified based on available observations. Hence, model calibration and selection, which are intrinsically linked, are essential steps of the workflow. Here, some procedures for model calibration and a criterion, the Akaike Information Criterion, of model selection based on experimental data are described. Rough derivation, practical technique of computation and use of this criterion are detailed.

1. Motivation

Richard Casement wrote (Casement, 1984) "It is a common fallacy to confuse scientists' models of reality with reality itself. A model is a map. A map is not the territory it describes". For instance, if we wanted to cross Canada by bike, we would need a road map of Canada with the elevation and inclination to be able to select the shortest route with the least difference in elevations. A political map of Canada representing the provinces, which is another model of Canada, will however be useless for the purpose of crossing Canada by bike. Hence, before writing a mathematical model, the question to be investigated has to be carefully defined, since a good answer to a poor question is still a poor answer (Burnham & Anderson, 2002).

Once the question is determined, the modelling process comprises several steps that can be organized in a recursive way. Based on experimental data available, the model variables and parameters are defined. To make the model as simple as possible, assumptions have to be made, identifying the important processes governing the problem investigated in the perspective of the question considered. Then, using basics principles governing the variables considered, such as physical laws or types of interactions, equations of the model can be written using the appropriate mathematical formalism. The well-posedness of the model such as the consistency of units, existence and uniqueness of solutions and non-negativity of solutions (if needed) have to be verified.

Once the model is written, its mathematical analysis has to be conducted to characterize its behaviour using the appropriate mathematical techniques and theories. Numerical experiments then need to be conducted. Adequate numerical

E-mail addresses: Stephanie.Portet@umanitoba.ca, stephanie.portet@umanitoba.ca.
Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2019.12.010

methods have to be carefully chosen to ensure the accuracy of numerical solutions of the model. Numerical experiments require the choice of values for model parameters. These values can be chosen from published work or estimated by calibrating the model responses to experimental data. Sensitivity analysis can help to understand the effects of model inputs (parameters or initial conditions) on model outputs and identify parameters that are the key drivers of the model responses. Finally, the model must represent accurately the observed process, it must reproduce known states of the real process. If several models are considered, model selection has to be used to identify the best model to represent the data. Ultimately, the proposed mechanism identified by the (best) model has to be validated by further experiments. The typical workflow used in mathematical biology is represented in Fig. 1.

The next example, adapted from (Zhang, Cao, & J Carroll, 2015), motivates the use of a collection of models and gives an illustration of its design.

Example 1.1. If we want to investigate the interactions between a population of rabbits and a population of foxes on a specific territory in which it is known that the foxes are specialist predators (foxes feed only on rabbits), we can propose a mathematical model using the formalism of ordinary differential equations. To do so, the following mechanisms have to be described:

- dynamics of predators in absence of preys,
- dynamics of preys in absence of predators,
- interactions between preys and predators.

For the dynamics of predators in absence of preys, the assumptions are pretty specific and allow us to write that the dynamics shows exponential decay. However, for the rabbit population, we have no specification. Hence, we could describe the dynamics of rabbits in absence of predators as an exponential growth, to translate an unlimited growth of the rabbit population, or as a logistic term to specify the limitation of the rabbit population due to the environmental supply. Similarly, for the interactions between the two populations, the predator-prey link can be described by different functional responses; a mass action type term or a saturating rate could be hypothesized. The saturating rate would describe the fact that even if the amount of rabbit is infinite, foxes cannot consume rabbits faster than some maximal rate. Different translations of the saturating process can be used; see for instance Fig. 2. Here, to write the model, we use a Michaelis-Menten type or Holling type II saturating term (left panel of Fig. 2).

The different assumptions yield different terms and then different models. Here, we consider three candidate models.

Model I	Model II	Model III
$\frac{dR}{dt} = aR - bRF$	$\frac{dR}{dt} = a\left(1 - \frac{R}{k}\right)R - bRF$	$\frac{dR}{dt} = aR - \frac{bRF}{1+fR}$
$\frac{dF}{dt} = -cF + eRF$	$\frac{dF}{dt} = -cF + eRF$	$\frac{dF}{dt} = -cF + \frac{eRF}{1+fR}$

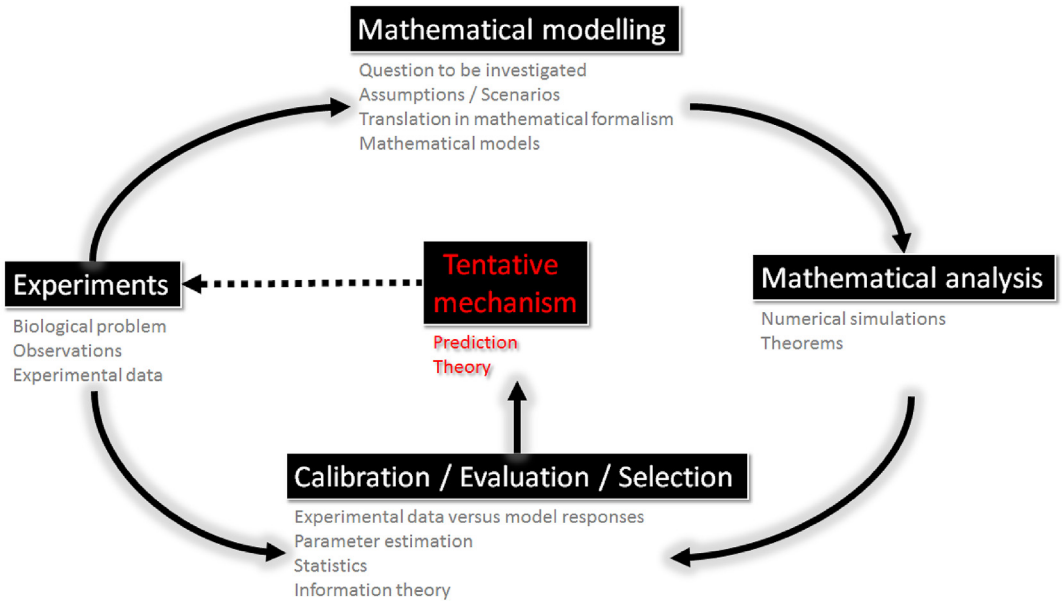

Fig. 1. Mathematical biology workflow; adapted from (Portet, 2015).

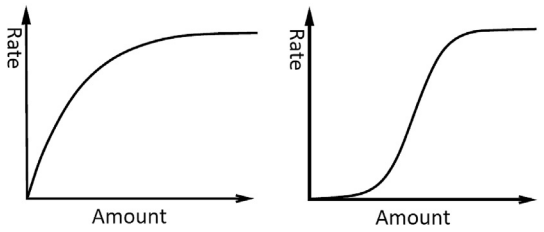

Fig. 2. Different translations of saturating rates. (Left) Hyperbolic saturation: as the amount increases the rate increases but slowing down (Michaelis-Menten dynamics or Holling type II function). (Right) Sigmoidal saturation: from a slow to rapid rate, "switch-like" rise toward to the limiting value or Holling type III function (Cooperativity).

Model I is the famous Lotka-Volterra model in which the rabbits grow exponentially. Model II describes a limitation of the rabbit population due to environmental factors. Model III is characterized by a saturating rate in the functional response. These three models are non-nested, as, for instance, Model II cannot be obtained from any other models. However, we can also define a general model

$$\frac{dR}{dt} = aR - vR^2 - \frac{bRF}{1+fR}$$
$$\frac{dF}{dt} = -cF + \frac{eRF}{1+fR}.$$

Each of the three models is a particular case of the general model obtained by setting some parameters to zero:

- $v = f = 0 \Rightarrow$ Model I (Lotka-Volterra model),
- $f = 0 \Rightarrow$ Model II (with logistic dynamics for preys),
- $v = 0 \Rightarrow$ Model III (with saturating rate in functional response).

The three models are therefore nested models of the general model. Note that from the general model, more nested models can be defined by setting other subsets of parameters to zero.

In Example 1.1, due to lack of knowledge of the processes, we had to consider different assumptions for a given mechanism. Furthermore, we could have used different translations for the saturating rate; that yields the definition of a collection of models. The design of a collection of models can be also used to mimic the well known positive and negative control experimental protocol used in experimental sciences. Considering a collection of models will allow the identification of a most plausible scenario for the defined problem. Systematic modelling of all possible scenarios is a powerful approach; see, for instance, (Gotoh et al., 2016).

Once we have a collection of models depending on unknown parameters, two questions have to be addressed. What are the parameter values? What is the best model to represent the experimental data? The first question will be solved using parameter estimation or model calibration methods. The second question can be addressed by model selection methods, which can be divided into two types:

- information theory criteria,
- statistical tests.

Here, only information theory based selection model methods are presented.

2. Model calibration

A mathematical model depends on independent variable(s), dependent variable(s) or state variable(s) and parameters p. Parameters of the model have (biological) interpretations and their values are unknown. Hence, we need to estimate the model parameters (find appropriate values) from measurements (the experimental data) in presence of errors in measurements. Here, two general (optimal) methods for parameter estimation are presented:

- Least Squares (LS), in which an objective function (sum of squared residuals of all measurements) is minimized.
- Maximum Likelihood (ML), in which the likelihood function is maximized.

Both LS and ML methods are optimization problems and provide point estimates. Other types of methods such as Bayesian inferences use the likelihood combined with priors on parameters to provide posterior distributions of parameters (Coelho, Codeço, & Gomes, 2011; Friston, 2002).

After presenting LS and ML methods, we will show that if the measurement errors are independent and normally distributed with a common variance, both methods are equivalent and give the same estimate for the mathematical model parameters.

2.1. A few words on optimization

When an analytic expression of the function to be optimized $f(p)$ is known and has appropriate properties, the following well known result can be used to find local extrema.

Theorem 2.1. *A smooth function $f(p)$ attains a local minimum (resp. maximum) at $\widehat{p}$ if the following two conditions hold*

- *the gradient $\frac{\partial f(p)}{\partial p}$ vanishes at $\widehat{p}$*
- *the Hessian $H(p)$ with (i,j)th entry $\frac{\partial^2 f(p)}{\partial p_i \partial p_j}$ is positive definite (resp. negative definite) at $\widehat{p}$, or*

$$z^T H(p) z > 0 \quad (resp. \ < 0),$$

where z is any real non-zero vector.

If $f(p)$ is non-smooth, the local extrema are at the discontinuity of $f(p)$ or where the gradient $\frac{\partial f(p)}{\partial p}$ is discontinuous or vanishes.

However, when the analytic expression of the function to be optimized $f(p)$ is unknown or lacks appropriate properties, search methods have to be used. These methods can be categorized as local or global optimization methods (Pitt & Banga, 2019; Sagar, LeCover, Shoemaker, & Varner, 2018). Local optimization methods are faster but might converge to local optima that are suboptimal solutions or only find a global optimum for appropriate starting points. Gradient descent-based methods such as Levenberg-Marquardt or Gauss-Newton or derivative-free local search methods such as simplex or Nelder-Mead are local optimization methods. Global optimization methods, which are heuristic or meta-heuristic methods, are more time consuming. Some examples of global optimization methods are the simulated annealing, genetic algorithms or particle swarm methods.

2.2. Least squares

The aim of the LS method is to find parameter values of the model that minimize the distance between data and model output. First, a simple case of the problem to solve is described.

- **Observation**: we have n data points (t_i, y_i) with $i = 1, \ldots, n$, where t_i are the values of the independent variable t.
- **Model**: the model is a differentiable function $f(t,p)$ that depends on t and p, the latter being a $k-$ vector of parameters (k parameters).
- **Criterion**: the total error between the model and data is defined as the residual sum of squares (*RSS*), the sum of squared residuals,

$$RSS(p) = \sum_{i=1}^{n} (y_i - f(t_i, p))^2 \tag{1}$$

and depends on p. The residual is the difference between the actual value y_i (data) and the predicted value $f(t_i, p)$ at a value of the independent variable t_i.

- **Solution**: the solution is the $k-$ vector of parameter values $\widehat{p}_{lsq}$ (the LS Estimate) that minimizes the sum of squares of vertical distances between data points and model points,

$$RSS\left(\widehat{p}_{lsq}\right) = \min_p RSS(p).$$

Hence, based on Theorem 2.1, possible solutions for $\widehat{p}_{lsq}$ can be obtained by setting the gradient equal to zero

$$\frac{\partial RSS}{\partial p_j} = 0, \quad j = 1, \ldots, k,$$

yielding

$$-2\sum_{i=1}^{n}(y_i - f(t_i, p))\frac{\partial f(t_i, p)}{\partial p_j} = 0, \quad j = 1, \ldots, k. \tag{2}$$

Then, solve the k equations for p_j with $j = 1, \ldots, k$.

Example 2.1. To illustrate, we consider the temporal evolution of the US population from 1790 to 2000; data are obtained from the US census (Fig. 3). Observations consist of $n = 22$ data points $(t_i, y_i) =$ (year, population) with $i = 1, \ldots, 22$. Based on the trend of data, we can hypothesize different forms for the model f, such as

- A quadratic function in t (years, the independent variable),

$$f(t, a, b, c) = y = at^2 + bt + c;$$

then $k = 3$ parameters are to be estimated and $p = (a, b, c)$.

- An exponential function in t (years),

$$f(t, a, b) = y = ae^{bt}.$$

Using the change of variable $\ln y = Y$, we obtain

$$\ln y = Y = \ln a + bt = A + bt.$$

Here, $k = 2$ parameters are to be estimated and $p = (A, b)$.

Notice that both models are linear in the parameters; hence, equation (2) to be solved to estimate parameter values is a linear system in p. Here, we only explicitate the solution of the problem when the exponential function is used.

When the exponential form is considered, the change of variable $\ln y = Y$ is used to make the model linear in the parameters; hence, the analytic expression of the RSS (1) takes the following form:

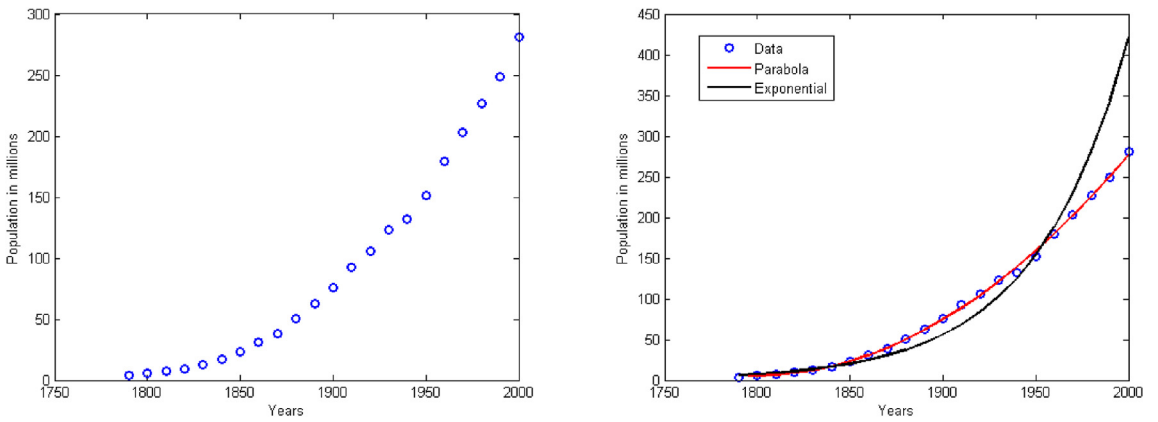

Fig. 3. (Left) US population (in millions) from 1790 to 2000; data from the US census, https://www.u-s-history.com/pages/h980.html. (Right) Data and best fits of the quadratic function $f(t) = 0.0067t^2 - 24.0358t + 21620.47$ and exponential function $f(t) = 1.2162 \times 10^{-15}e^{0.0202t}$.

$$RSS(A,b) = \sum_{i=1}^{n} (\ln y_i - (A + bt_i))^2.$$

We want to find values of A and b that minimize $RSS(A,b)$. By setting its gradient to zero to find critical points of $RSS(A,b)$, (2) is then

$$\sum_{i=1}^{n} (\ln y_i - (A + bt_i)) \frac{\partial (A + bt_i)}{\partial A} = \sum_{i=1}^{n} (\ln y_i - (A + bt_i)) = 0,$$
$$\sum_{i=1}^{n} (\ln y_i - (A + bt_i)) \frac{\partial (A + bt_i)}{\partial b} = \sum_{i=1}^{n} (\ln y_i - (A + bt_i)) t_i = 0.$$

These two necessary conditions for the existence of a local extremum can be rewritten in matrix form as follows

$$\begin{bmatrix} n & \sum_{i=1}^{n} t_i \\ \sum_{i=1}^{n} t_i & \sum_{i=1}^{n} t_i^2 \end{bmatrix} \begin{bmatrix} \widehat{A} \\ \widehat{b} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \ln y_i \\ \sum_{i=1}^{n} t_i \ln y_i \end{bmatrix},$$

where $\widehat{A}$ and $\widehat{b}$ (LS estimates of A and b) are the unique solutions if the matrix is non-singular. In such a case, using Cramer's rule, the unique solution that minimizes RSS is

$$\widehat{A} = \frac{\sum_{i=1}^{n} \ln y_i \sum_{i=1}^{n} t_i^2 - \sum_{i=1}^{n} t_i \ln y_i \sum_{i=1}^{n} t_i}{n \sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i \right)^2},$$

$$\widehat{b} = \frac{n \sum_{i=1}^{n} t_i \ln y_i - \sum_{i=1}^{n} \ln y_i \sum_{i=1}^{n} t_i}{n \sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i \right)^2}.$$

Results are given on the right panel of Fig. 3.

Note that the two models considered in Example 2.1 are not explanatory models but descriptive models. They have a strong predictive power but they cannot encode any underlying process to explain the trend in data. For instance, with these two models, we cannot decide if the increase of the population is due to either an increase in fertility, a decrease in the death rate or the contribution of immigration. To unravel the processes responsible for the trend in data, a mechanistic approach needs to be followed and explanatory models need to be designed by modelling the hypothesized underlying mechanisms such as, for instance, in Example 1.1. Hence, when the ODE formalism is used, explanatory models can be expressed as follows

$$\frac{dx}{dt} = m(x,p,t),\ x(t_0) = x_0(p), \qquad \text{with} \quad h(x,p,t) = \tilde{y}, \tag{3}$$

where t is the independent variable, $x(t)$ is the vector of state variables, x_0 is the vector of initial conditions and p is the vector of unknown constant parameters. The observable function h depends on model inputs and outputs.

Now, the LS method detailed above is generalized. To find the vector of parameter values p that minimizes the distance between measured and simulated observations, the scalar objective function (cost function), $F_{ls}(p)$, is defined:

$$F_{ls}(p) = \sum_{e=1}^{n_e} \sum_{o=1}^{n_o^e} \sum_{i=1}^{n_i^{e,o}} \omega_i^{e,o} \left(y_e^o(t_i) - \tilde{y}_o^e(t_i,p) \right)^2, \tag{4}$$

with n_e the number of experiments, n_o^e the number of observables per experiments and $n_i^{e,o}$ the number of samples per observable per experiment. Also, $y_e^o(t_i)$ are measured data, $\omega_i^{e,o}$ are weights and $\tilde{y}_o^e(t_i,p)$ are simulated observable outputs. Then, an optimization method to minimize $F_{ls}(p)$ has to be chosen to find $\widehat{p}_{LSE}$ (the LS Estimate) such that

$$F_{ls}(\widehat{p}_{LSE}) = \min_p F_{ls}(p).$$

Note that when models as defined in (3) are considered, analytic expressions of observables $\tilde{y}$ might not exist; in this case, only their numerical approximations can be used to compute (4). Hence, prior to any model calibration, the mathematical

analysis of the ODE model (3) has to be carried out to characterize the model behaviour in as much detail as possible. For instance, if the model (3) presents an oscillatory regime, specific search methods will have to be used (Pitt & Banga, 2019).

2.3. Maximum likelihood

Experimental data $y = (y_1, \ldots, y_n)$ can be interpreted as a random sample generated from an unknown probability distribution function (pdf) depending on parameters $p = (p_1, \ldots, p_k)$. Then, a model is defined as a family of probability distributions indexed by the model's parameters,

$$f(y|p) = \text{probability of observing data } y \text{ given the parameter } p.$$

If the observations y_i are statistically independent of one another, the pdf of observing the data $y = (y_1, \ldots, y_n)$ given the parameter vector p is the multiplication of the pdfs for individual observations

$$f((y_1, \ldots, y_n)|(p_1, \ldots, p_k)) = f(y_1|(p_1, \ldots, p_k))f(y_2|(p_1, \ldots, p_k))$$
$$\cdots f(y_n|(p_1, \ldots, p_k)).$$

The function $f((y_1, \ldots, y_n)|(p_1, \ldots, p_k))$ is the probability of observing data y for a given value of p and is a function of data y. Varying the parameter p across its range defines a family of pdfs that makes up the model.

Hence, the inverse problem to solve is to find the pdf among all the pdfs of the family that is the most "likely" to have produced the data y. To attack the problem, a function of parameters p needs to be defined. Hence, the likelihood function $\mathscr{L}$ is the density function regarded as a function of p

$$\mathscr{L}(p|y) = f(y|p).$$

The role of data y and parameters p is reversed. The likelihood of a particular value for a parameter is the probability of obtaining the observed data y if the parameter had that value. It measures how well the data supports that particular value of parameters. The density function f (function of y, data scale) gives the probability of observing y given the parameter p and sums to 1 over all the possible values of y. The likelihood function $\mathscr{L}$ is a function of p (parameter scale) given the data and does NOT sum to 1 over the possible values of p.

The inverse problem to solve is equivalent to seeking the value $\widehat{p}_{MLE} = (\widehat{p}_{1,MLE}, \ldots, \widehat{p}_{k,MLE})$ of the parameter vector p that maximizes the likelihood function $\mathscr{L}(p|y)$. In practice, the log-likelihood $\ln\mathscr{L}(p|y)$ will be considered. As the logarithm is a monotonically increasing function, maximizing $\mathscr{L}(p|y)$ is equivalent to maximizing the log-likelihood $\ln\mathscr{L}(p|y)$ to find the Maximum Likelihood Estimator $\widehat{p}_{MLE}$. A $\widehat{p}_{MLE}$ satisfies the following conditions based on Theorem 2.1:

- Necessary condition of existence of a $\widehat{p}_{MLE}$,

$$\frac{\partial \ln\mathscr{L}(p|y)}{\partial p_i} = 0, \quad i = 1, \ldots, k.$$

- Convexity condition: consider the Hessian matrix $H(p)$, $H_{i,j}(p) = \frac{\partial^2 \ln\mathscr{L}(p|y)}{\partial p_i \partial p_j}$ with $i,j \in \{1, \ldots, k\}$. Then there must hold that

$$z^T H(p) z < 0,$$

where z is any real non-zero $k-$ vector.

Example 2.2. Suppose the blood pressure of 1000 patients is recorded with observations given in gray in Fig. 4. Seeing the shape of the data in Fig. 4, we assume that experimental data $y_1, \ldots, y_n$ (y_i is the blood pressure of the patient i with $i \in \{1, \ldots 1000\}$) are drawn from a normal distribution $\mathscr{N}(\mu, \sigma^2)$ with μ and σ unknown ($k = 2$, $p = (\mu, \sigma)$).

Let $Y_1, \ldots, Y_n$ be n i.i.d.[1] $\mathscr{N}(\mu, \sigma^2)$ random variables,

[1] independent and identically distributed.

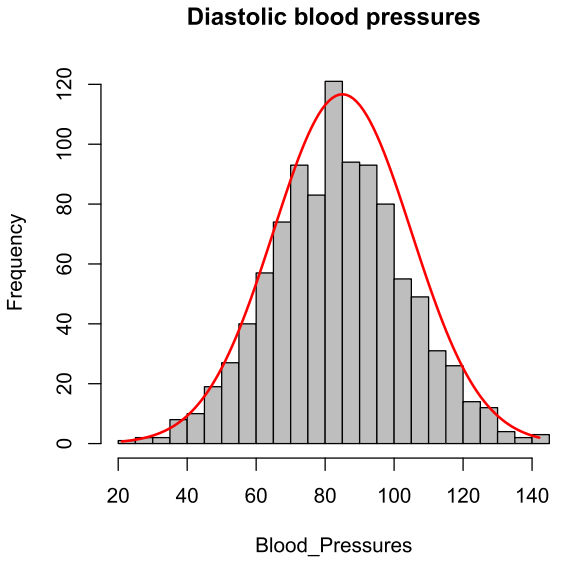

Fig. 4. Blood pressures of 1000 patients (data generated in R). In gray, observations that looks normally distributed. In red, the normal distribution $\mathcal{N}(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE}) = \mathcal{N}(85, 20^2)$.

$$f_{Y_i}(y_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}.$$

The joint probability distribution function is

$$f(y_1,\ldots,y_n|\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(y_i-\mu)^2}{2\sigma^2}}.$$

Hence, for a fixed data set $y_1,\ldots,y_n$, the likelihood function is

$$\mathscr{L}(\mu,\sigma|y_1,\ldots,y_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(y_i-\mu)^2}{2\sigma^2}}$$

and the log-likelihood function is

$$\ln\mathscr{L}(\mu,\sigma|y_1,\ldots,y_n) = -n\left(\ln\left(\sqrt{2\pi}\right) + \ln(\sigma)\right) - \sum_{i=1}^n \frac{(y_i-\mu)^2}{2\sigma^2}.$$

To find the Maximum Likelihood Estimate of μ, $\widehat{\mu}_{MLE}$, we differentiate the log-likelihood function with respect to μ,

$$\frac{\partial \ln\mathscr{L}(p|y)}{\partial\mu} = \sum_{i=1}^n \frac{(y_i-\mu)}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^n y_i = n\widehat{\mu}_{MLE} \Rightarrow \widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}.$$

To find the Maximum Likelihood Estimate of σ, $\widehat{\sigma}_{MLE}$, we differentiate the log-likelihood function with respect to σ,

$$\frac{\partial \ln\mathscr{L}(p|y)}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(y_i-\mu)^2}{\sigma^3} = 0 \Rightarrow \widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (y_i-\mu)^2.$$

Using $\mu = \widehat{\mu}_{MLE}$ we obtain $\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{\mu}_{MLE})^2$. Maximum Likelihood Estimates are $\widehat{\mu}_{MLE}$ (the mean of data) and $\widehat{\sigma}^2_{MLE}$ (the variance of data). Using the data, we compute $\widehat{\mu}_{MLE} = 85$ and $\widehat{\sigma}^2_{MLE} = 20^2$ and we plot in red the normal distribution function $\mathcal{N}(85, 20^2)$ in Fig. 4. Furthermore, the maximum log-likelihood can be computed by replacing μ and σ by their estimates as follows:

$$\ln\mathscr{L}(\widehat{\mu}_{MLE}, \widehat{\sigma}_{MLE}|y_1, \ldots, y_n) = \max_{\mu,\sigma} \ln\mathscr{L}(\mu, \sigma|y_1, \ldots, y_n).$$

For many problems, even computational evaluation of the likelihood is infeasible. A relationship between the negative of the log-likelihood function and the residual sum of squares (or the cost function used in least squares method) is now derived under some condition.

2.4. Relationship between least squares estimate and maximum likelihood estimate

Define the statistical model

$$Y_e^o = \tilde{y}_o^e(t_i, p) + \varepsilon_e^o,$$

where the measured data $y_e^o(t_i)$ is a realization of the random variable Y_e^o, $\tilde{y}_o^e(t_i, p)$ is the (algebraic, ODE or PDE) mathematical model output and the random variable ε_e^o represents measurement errors or noise (Miao, Dykes, Demeter, & Wu, 2009; Zhang et al., 2015). Assuming independent and normally distributed additive measurement errors with standard deviation $\sigma_i^{e,o}$, the probability of observing the data y given the parameters p and $\sigma_i^{e,o}$ is

$$f(y|\theta) = \prod_{e=1}^{n_e}\prod_{o=1}^{n_o^e}\prod_{i=1}^{n_i^{e,o}} \frac{1}{\sqrt{2\pi}\sigma_i^{e,o}} exp\left(-\frac{1}{2}\left(\frac{y_e^o(t_i) - \tilde{y}_o^e(t_i, p)}{\sigma_i^{e,o}}\right)^2\right),$$

where $y_e^o(t_i)$ are measured data, $\tilde{y}_o^e(t_i, p)$ model output and θ includes the mathematical model parameters p and the statistical model parameters $\sigma_i^{e,o}$. As previously defined, n_e is the number of experiments, n_o^e is the number of observables per experiment and $n_i^{e,o}$ is the number of samples per observable per experiments. Hence, the likelihood function is

$$\mathscr{L}(\theta|y) = \prod_{e=1}^{n_e}\prod_{o=1}^{n_o^e}\prod_{i=1}^{n_i^{e,o}} \frac{1}{\sqrt{2\pi}\sigma_i^{e,o}} exp\left(-\frac{1}{2}\left(\frac{y_e^o(t_i) - \tilde{y}_o^e(t_i, p)}{\sigma_i^{e,o}}\right)^2\right).$$

Maximizing the likelihood is equivalent to minimizing the negative of the log-likelihood function, so we consider the negative of the log-likelihood function, defined as

$$-\ln\mathscr{L}(\theta|y) = \frac{1}{2}\sum_{e=1}^{n_e}\sum_{o=1}^{n_o^e}\sum_{i=1}^{n_i^{e,o}}\left[\ln\left(2\pi\left(\sigma_i^{e,o}\right)^2\right) + \left(\frac{y_e^o(t_i) - \tilde{y}_o^e(t_i, p)}{\sigma_i^{e,o}}\right)^2\right].$$

Recall the cost function, $F_{ls}(p)$, defined in the least squares method by (4). Hence, there is the following relationships between the negative log-likelihood $-\ln\mathscr{L}(\theta|y)$ and least squares cost function $F_{ls}(p)$ (Baker, Bocharov, Paul, & Rihan, 2005).

- For $\omega_i^{e,o} = 1/(\sigma_i^{e,o})^2$ (weighted LS),

$$-\ln\mathscr{L}(\theta|y) = \frac{1}{2}\sum_{e=1}^{n_e}\sum_{o=1}^{n_o^e}\sum_{i=1}^{n_i^{e,o}}\ln\left(2\pi\left(\sigma_i^{e,o}\right)^2\right) + \frac{1}{2}F_{ls}(p) = n_e n_o^e n_i^{e,o}\ln\left(\sqrt{2\pi}\right) + \ln\left(\prod_{e=1}^{n_e}\prod_{o=1}^{n_o^e}\prod_{i=1}^{n_i^{e,o}}\sigma_i^{e,o}\right) + \frac{1}{2}F_{ls}(p).$$

Hence, $-\ln\mathscr{L}(\theta|y)$ and $F_{ls}(p)$ have the same optimal mathematical model parameters p ($\widehat{p} = \widehat{p}_{MLE} = \widehat{p}_{LSE}$).

- For $\sigma_i^{e,o} = \sigma$ (weighted LS) and $\omega_i^{e,o} = 1$ (ordinary LS),

$$-\ln\mathscr{L}(\theta|y) = n_e n_o^e n_i^{e,o}\left(\ln\left(\sqrt{2\pi}\right) + \ln(\sigma)\right) + \frac{1}{2\sigma^2}F_{ls}(p).$$

Hence, $-\ln\mathscr{L}(\theta|y)$ and $F_{ls}(p)$ have the same optimal parameters for the mathematical model parameters p ($\widehat{p} = \widehat{p}_{MLE} = \widehat{p}_{LSE}$) and from $\frac{\partial\ln\mathscr{L}(\theta|y)}{\partial\sigma} = 0$,

$$\widehat{\sigma}^2 = \frac{1}{n_e n_o^e n_i^{e,o}} F_{ls}(\widehat{p}).$$

Thus, the minimum of the negative of the log-likelihood function is

$$-\ln \mathscr{L}(\widehat{\theta}_{MLE}|y) = \frac{n_e n_o^e n_i^{e,o}}{2}\ln(2\pi) + \frac{n_e n_o^e n_i^{e,o}}{2} + \frac{n_e n_o^e n_i^{e,o}}{2}\ln\left(\underbrace{\frac{F_{ls}(\widehat{p})}{n_e n_o^e n_i^{e,o}}}_{\text{MLE of variance}}\right), \tag{5}$$

where $\widehat{p} = \widehat{p}_{MLE} = \widehat{p}_{LSE}$.

Therefore, under the assumption of independent, normally distributed additive measurement errors with constant variance, the Least Squares Estimates and Maximum Likelihood Estimates of the mathematical model parameters are the same.

2.5. Problems in model calibration

The use of optimal methods for model calibration (LS or ML) for non-linear models have inherent problems that can be categorized as follows.

- Lack of prior knowledge about parameters.
- Lack of identifiability (non-uniqueness of optimal solutions for parameter estimation) (Miao et al.,2009). There exist two types of identifiability:
  1 (Practical identifiability) The larger the number of unknown parameters in a model, the larger the amount of quantitative data necessary to determine meaningful values for these parameters.
  2 (Structural identifiability) Even if appropriate experimental data are available, model parameters may not be uniquely identifiable.
- Convergence to local optima (ill-conditioning and non-convexity).
- Overfitting (fitting the noise instead of the signal).

Example 2.3. A famous example illustrates the problem of structural identifiability (Janzén et al., 2016; Yates, 2006). A two compartment model describing the temporal evolution of the amount of drug injected in the blood system is considered (Fig. 5):

$$\begin{aligned}
\frac{dx_1}{dt} &= -(k_{12}+k_{10})x_1 + k_{21}x_2, \qquad x_1(0) = D,\\
\frac{dx_2}{dt} &= k_{12}x_1 - (k_{21}+k_{20})x_2, \qquad x_2(0) = 0,\\
\tilde{y} &= \frac{x_1}{V_1},
\end{aligned}$$

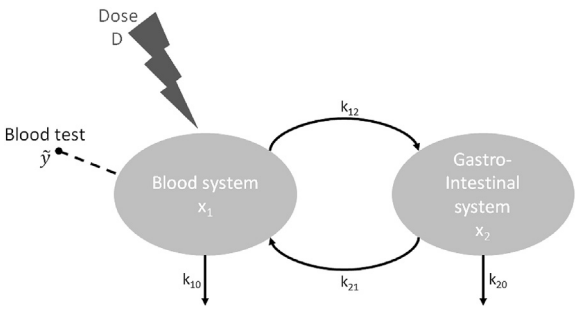

Fig. 5. Pharmacokinetics model: evolution of the amount of drug in the blood system x_1, the gastro-intestinal system x_2 and the blood test, which is the only observable outputs of the model.

where x_1 (variable) is mass of drug in compartment 1 (blood system), x_2 (variable) mass of drug in compartment 2 (gastro-intestinal system), $\tilde{y}$ (variable) is the observable concentration of drug in the blood test with V_1 (parameter) the volume of the observed compartment (blood system). The parameters k_{ij} are rates of transfer from i to j. The dose D of drug injected in the blood system is a known parameter. Hence, there are five parameters to estimate.

The model is a homogeneous linear system that can be solved explicitly; closed forms of $x_1(t)$ and $x_2(t)$ can be obtained. Hence, the time solution $\tilde{y}(t)$ can be expressed as follows:

$$\tilde{y}(t) = C_1 e^{-\lambda_1 t} + C_2 e^{-\lambda_2 t}.$$

As seen in Example 2.1, functions of the form $C_i e^{-\lambda_i t}$ can be linearised by a change of variables; then, C_1, C_2, λ_1, and λ_2 can be determined from the observed concentration curve in a unique way. However, their knowledge results in only four conditions/equations that depend on the five unknown parameters k_{ij} and V_1. Hence, we cannot find a unique solution for the five parameters to estimate. Therefore, the model is structurally unidentifiable.

3. Model selection

When a collection of models is considered to investigate a problem, after calibrating each model to the available experimental data, a model selection method can be employed to discriminate between the different candidate models of the collection in terms of the representation of experimental data considered.

3.1. Naive approach

For linear models, a naive approach to compare models is to used the R^2 or adjusted R^2. They are a measure of the goodness of fit; the best fit is selected but the model complexity is neglected. The R^2 is defined as

$$R^2 = 1 - \frac{RSS/n}{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}$$

or by replacing the two variances with their unbiased estimates, the adjusted R^2 is obtained,

$$R^2_{adj} = 1 - \frac{RSS/(n-k-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)},$$

where RSS is the residual sum of squares as defined in (1), n is the sample size, y are the data, $\bar{y}$ is the average of the data and k is the number of parameters. These criteria select the model that maximizes R^2 or R^2_{adj}; the most parameter rich model is selected.

However, in selecting the best model, the principle of parsimony, which states that a model should be as simple as possible, should be employed (Fig. 6). A good model is a proper balance between underfitting and overfitting. Johnson and Omland (Johnson & Omland, 2004) state the following: "Parsimony is, in statistics, a trade-off between bias and variance. Too few parameters results in high bias in parameter estimators and an underfit model (relative to the best model) that fails to identify all factors of importance. Too many parameters results in high variance in parameter estimators and an overfit model that risks identifying spurious factors as important, and that cannot be generalized beyond the observed sample data".

Here a model selection method is introduced, the Akaike Information Criterion, which accounts for the goodness of the fit and parsimony principle.

3.2. Sketch of Akaike Information Criterion derivation

First we need to define the Kullback-Leibler (KL) divergence, which is used to measure the difference between two probability distributions $f(x)$ and $g(x)$ defined over the same probability space (Kullback & Leibler, 1951). The KL divergence of $g(x)$ from $f(x)$ is

$$I(f,g) = \sum_{x \in X} f(x) \ln\left(\frac{f(x)}{g(x)}\right),$$

if $f(x)$ and $g(x)$ are pdfs of a discrete random variable X and

$$I(f,g) = \int_X f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx,$$

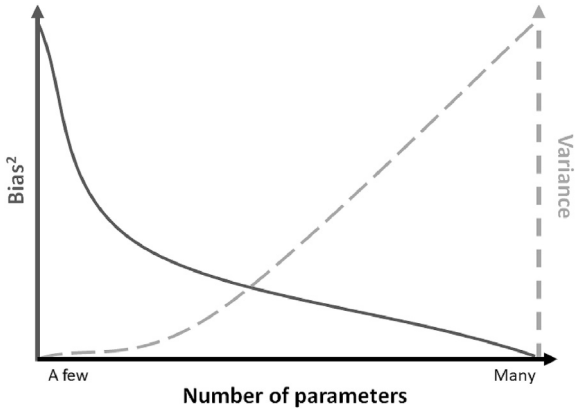

Fig. 6. Principle of parsimony.

if $f(x)$ and $g(x)$ are pdfs of a continuous random variable X.

The KL divergence can be interpreted as the distance between two probability distributions; however, it is not really a distance as the KL divergence is not symmetrical. Here are some properties of the KL divergence:

- $I(f,g) \neq I(g,f)$ (not symmetric),
- $I(f,g) \geq 0$,
- $I(f,g) = 0$ if and only if $f = g$.

In 1973, Akaike (AkaikePetrov and Csaki, 1973) found a relationship between the maximum likelihood (statistical analysis) and Kullback-Leibler divergence (information theory) and, based on that, defined a model selection criterion, now called Akaike Information Criterion (AIC). A sketch of the AIC derivation is given here.

As no single model includes the whole truth or the complete information about the phenomena under investigation, the quantity of information loss has to be determined. The Kullback-Leibler divergence can be used as a measure to quantify the information lost when approximating the full reality $f(x)$ by a model $g(x|\theta)$, where the model $g(x|\theta)$ depends on parameters θ,

$$I(f,g) = \int f(x) \ln\left(\frac{f(x)}{g(x|\theta)}\right) dx.$$

However, the reality/truth f is unknown and the parameters θ must be estimated from data y (generated from f).

Measurement of the information lost when approximating the full reality $f(x)$ by a model $g(x|\theta)$, $I(f,g)$, can be rewritten as the difference between two expectations with respect to the true distribution f as follows:

$$\begin{aligned} I(f,g) &= \int f(x) \ln\left(\frac{f(x)}{g(x|\theta)}\right) dx \\ &= \int f(x) \ln(f(x)) dx - \int f(x) \ln(g(x|\theta)) dx \\ &= E_f[\ln(f(x))] - E_f[\ln(g(x|\theta))], \end{aligned}$$

where the first term $E_f[\ln(f(x))] = C$, which depends only on the unknown true distribution f, is unknown and constant. The second term is the relative KL divergence between f and g and is defined as

$$I(f,g) - C = -E_f[\ln(g(x|\theta))].$$

Consider two models g_1 and g_2. If $I(f,g_1) < I(f,g_2)$ then model g_1 is better than model g_2. In other words,

$$\begin{aligned} I(f,g_1) &< I(f,g_2) \\ I(f,g_1) - C &< I(f,g_2) - C \\ -E_f\Big[\ln(g_1(x|\theta))] &< -E_f[\ln(g_2(x|\theta))\Big]. \end{aligned}$$

Thus, without knowing C, we know how much better g_1 is than g_2 (comparison between 2 models),

$$I(f,g_2)-I(f,g_1)=-E_f\left[\ln(g_2(x|\theta))\right]+E_f[\ln(g_1(x|\theta))\Big].$$

Hence, the use of the relative KL divergences between f and the models g_i (instead of the KL divergence between f and g_i that compares the model g_i to the unknown truth f) allows the comparison of the candidate models g_i.

The models depend on parameters that have to be estimated using the data y. Let us define $\widehat{\theta}(y)$, the estimator of θ, as a random variable. Consequently, $I(f,g(\cdot|\widehat{\theta}(y)))$ is also a random variable. Hence, its expectation can be estimated,

$$E_y\left[I(f,g(\cdot|\widehat{\theta}(y)))\right]=C-E_y[E_x[\ln(g(x|\widehat{\theta}(y)))]],$$

where x and y are two independent random samples from the same distribution f and both statistical expectations are taken with respect to the truth f.

As we want to minimize the information loss by approximating the full reality f by models, we want to minimize the estimated expected KL divergence $E_y[I(f,g(\cdot|\widehat{\theta}(y)))]$ over the set of models considered. That is equivalent to maximizing the estimated expected relative KL divergence. Hence, a model selection criterion can be defined as follows:

$$\max_{g\in G}E_y[E_x[\ln(g(x|\widehat{\theta}(y)))]],$$

where G is the collection of models in terms of probability density functions.

In 1973, Akaike (AkaikePetrov and Csaki, 1973) found an asymptotically (for a large sample) unbiased estimator of the expected relative Kullback-Leibler divergence $E_y[E_x[\ln(g(x|\widehat{\theta}(y)))]]$, given by

$$\ln\mathscr{L}(\widehat{\theta}_{MLE}|y)-K,$$

where $\mathscr{L}$ is the likelihood function (previously defined in the section on model calibration), $\widehat{\theta}_{MLE}$ is the maximum likelihood estimate of θ ($\ln\mathscr{L}(\widehat{\theta}_{MLE}|y)$ is the maximum log-likelihood value) and K is the number of estimated parameters including the variance (the bias correction term). Akaike multiplied this estimator by -2. Thus, the Akaike Information Criterion for each model considered with the same data set is defined as

$$AIC=-2\ln(\mathscr{L}(\widehat{\theta}_{MLE}|y))+2K. \tag{6}$$

Therefore, the best model within the collection of models considered given the data is the one with the minimum AIC value. More details on the derivation of AIC can be found in (Burnham & Anderson, 2002).

3.3. Model selection using AIC

Consider a collection of R models. Which model of the collection would best represent reality given the data we have recorded? To answer this question, we can compute the information criterion for each model of the collection.

If the number of observations is large enough, $K<(N/40)$, use AIC

$$AIC=-2\ln(\mathscr{L}(\widehat{\theta}_{MLE}|y))+2K.$$

For a small number of observations, $K>(N/40)$, Sigiura (Hurvich & Tsai, 1989; Sugiura, 1978) developed the corrected AIC (AICc),

$$AICc=-2\ln(\mathscr{L}(\widehat{\theta}_{MLE}|y))+\frac{2KN}{N-K-1}=AIC+\frac{2K(K+1)}{N-K-1},$$

where K is the number of estimated parameters in the mathematical and statistical model and N is the number of observations. As $N\to\infty$, $AICc\to AIC$. Other extensions of the AIC have been derived to accommodate other specific cases (Burnham & Anderson, 2002).

In practice, except for some simple models, computing the minimum of the negative of the log-likelihood $-\ln(\mathscr{L}(\widehat{\theta}_{MLE}|y))$ for a model is difficult, in particular when dealing with an ODE or PDE model. However, under some assumptions, we have previously derived a relationship between the negative of the log-likelihood and least squares cost function, which can be used to easily compute AIC and AICc.

Recall that when the measurement errors are independent, identically and normally distributed with the same variance, the minimum of the negative of the log-likelihood function can be expressed as follows (see (5)),

$$-\ln\mathscr{L}(\widehat{\theta}_{MLE}|y) = \frac{N}{2}\ln(2\pi) + \frac{N}{2} + \frac{N}{2}\ln\left(\frac{F_{ls}(\widehat{p})}{N}\right),$$

with $\widehat{p} = \widehat{p}_{MLE} = \widehat{p}_{LSE}$ and the number of observations $N = n_e n_o^e n_i^{e,o}$. Hence, under this assumption and as the AIC or AICc are used to compare the models with each other and that the same data are used to estimate the parameters for all the models, the first term of (6) can be expressed as $-2\ln(\mathscr{L}(\widehat{\theta}_{MLE}|y)) \approx N\ln\left(\frac{F_{ls}(p)}{N}\right)$. Therefore, *AIC* and *AICs* can be computed as follows:

$$AIC = N\ln\left(\frac{F_{ls}(\widehat{p})}{N}\right) + 2K = N\ln\left(\frac{RSS}{N}\right) + 2K$$

and

$$AICc = N\ln\left(\frac{F_{ls}(\widehat{p})}{N}\right) + \frac{2KN}{N-K-1} = N\ln\left(\frac{RSS}{N}\right) + \frac{2KN}{N-K-1},$$

where K is the number of estimated parameters (numbers of mathematical model estimated parameters + 1 for the variance) and N is the number of observations.

Due to this approximation, interpreting the actual values of AIC and AICc has no real meaning. Furthermore, as only the estimates of the expected relative KL divergences between f and $g_i(x|\theta)$ are known with the information criteria, it is convenient to scale them with respect to the minimum AIC value among all models. Compute AIC_i of each model i with $i \in \{1, \ldots, R\}$ and then compute the AIC differences Δ_i as follows (Akaike, 1974):

$$\Delta_i = AIC_i - \min_i AIC_i,$$

where $\min_i AIC_i$ is the AIC value of the best model in the collection. The AIC difference Δ_i estimates the information loss when using model i rather than the estimated best model. Hence, the larger Δ_i, the less plausible is model i.

Some guidelines for the interpretation of AIC difference Δ_i in the case of nested models are given in (Burnham & Anderson, 2002):

- $\Delta_i \in [1, 2]$, model i has substantial support and should be considered,
- $\Delta_i \in [4, \ldots, 7]$, model i has less support,
- $\Delta_i > 10$, model i has no support and can be omitted.

However, these rules might be different for non-nested models or when a very large number of models is present.

Going further, for an easier interpretation, Δ_i can be rescaled. The likelihood of model g_i given the data y can be defined as

$$\mathscr{L}(g_i|y) \propto \exp\left(-\frac{\Delta_i}{2}\right).$$

The model likelihoods can be normalized so that they sum to 1. This normalization yields the definition of the Akaike weight or "weight of evidence" of model i for being the best model (in terms of KL) of the collection given the data recorded,

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R}\exp(-\Delta_r/2)}. \tag{7}$$

The Akaike weight w_i of model i can be interpreted as the probability that model i is the best (approximating) model given the experimental data and the collection of models considered. Hence, the smaller the weight w_i, the less plausible is model i. We can consider a single best model i if $w_i > 0.9$.

Moreover, by using Akaike weights, the evidence ratio of model i versus model j can be defined as

$$\frac{w_i}{w_j} = \frac{\mathscr{L}(g_i|y)}{\mathscr{L}\left(g_j\middle|y\right)}.$$

This evidence ratio quantifies the strength of evidence in favour of model i over model j.

Furthermore, using Akaike weights, the confidence set of models can be also defined. Rank models in order of Akaike weights (from the largest to smallest); then, compute the cumulative sum of their weights. The minimal subset of models whose the cumulative sum is larger than 0.95 constitutes the 95% confidence set of models (in terms of KL) (Symonds & Moussalli, 2011): there is 95% confidence that the best approximating model is in this subset.

Last but not least, Akaike weights can also be used to determine the relative importance of a process. The sum of Akaike weights over all models in which the process of interest appears gives a measure of the relative importance of the process of interest.

If the aim of the study is parameter estimation and when several models are well supported by the experimental data or when there is no model i with $w_i > 0.9$, we can use model averaging. If θ is a parameter common to all models well supported by the data, the weighted average of parameter estimate $\widehat{\theta}$ is

$$\widehat{\overline{\theta}} = \sum_{i=1}^{R} w_i \widehat{\theta}_i,$$

where $\widehat{\theta}_i$ is the MLE $\widehat{\theta}_{MLE}$ from model i. If θ is only common to a subset of models, rescale the w_i of the subset to have $\sum w_i = 1$.

Example 3.1. In Jacquier et al. (2018) (Jacquier et al., 2018), using a collection of mathematical models, the regulation of NMT1 by mTOR is investigated. mTOR is a kinase (enzyme that phosphorylates other molecules) that regulates cell growth, proliferation, survival and migration. mTOR exists in two forms: the inactive form (mTOR) and the active form (pmTOR). Cancer cells exploit mTOR to enhance their capacity to grow. Rapamycin is a drug used in cancer therapy that targets mTOR and prevents its activation by preventing mTOR phosphorylation. Strong expression of NMT1 has been reported in malignant breast tissues compared with normal breast cells.

Experimental data used in (Jacquier et al., 2018) shows that Rapamycin treatment decreases phosphorylation of mTOR (pmTOR) and augments total NMT1 levels over time. However, there is no significant change in the total mTOR levels (mTOR + pmTOR). Four dataset are recorded for this study. For each of the four datasets, three quantities are measured at $m = 9$ time points: total mTOR (T^{exp}_{total}), pmTOR (T^{exp}_{p}) and total NMT1 (N^{exp}_{total}).

The core assumption of the work is that the NMT1 phosphorylation is regulated by pmTOR; the general model described in the framed diagram of Fig. 7 is

$$\begin{aligned}
\frac{dT}{dt} &= \overbrace{\frac{-\alpha_T T}{K_T + T}}^{\text{phosphorylation}} + \overbrace{\frac{\alpha_{T_p} T_p}{K_{T_p} + T_p}}^{\text{dephosphorylation}} + \overbrace{\Pi}^{\text{synthesis}}{}_T - \overbrace{\beta T N}^{\text{feedback}} + \overbrace{g(T, R_c)}^{\text{rapamycin effect}}, \frac{dT_p}{dt} \\
&= \overbrace{\frac{\alpha_T T}{K_T + T}}^{\text{phosphorylation}} - \overbrace{\frac{\alpha_{T_p} T_p}{K_{T_p} + T_p}}^{\text{dephosphorylation}} - \overbrace{\delta_{T_p} T_p}^{\text{degradation}}, \frac{dN}{dt} = -\overbrace{\frac{\alpha_N T_p N}{K_N + N}}^{\text{phosphorylation}} + \overbrace{\Pi_N}^{\text{synthesis}}, \frac{dN_p}{dt} \\
&= \overbrace{\frac{\alpha_N T_p N}{K_N + N}}^{\text{phosphorylation}} - \overbrace{\delta_{N_p} N_p}^{\text{degradation}}, \frac{dR_c}{dt} = \overbrace{h(T, R_c)}^{\text{rapamycin effect}},
\end{aligned}$$

where T is the inactive form of mTOR, T_p is (pmTOR) the active form of mTOR, N is the active form of NMT1, N_p is the phosphorylated form of NMT1 (pNMT1) and R_c is the complex formed by Rapamycin and mTOR. In presence of Rapamycin, five models are considered to test alternative hypotheses difficult to test experimentally.

- Does the regulation of endogenous levels of mTOR components impact the dynamics? Two cases are described:
  - synthesis and degradation of the relevant forms of mTOR,
  - the total mTOR (mTOR + pmTOR) is assumed to be constant.
- Does NMT1 have a negative feedback effect on mTOR? Two cases are considered:
  - no feedback, $\beta = 0$,
  - presence of feedback, $\beta > 0$.
- Is the effect of rapamycin on mTOR reversible or irreversible? Two cases are considered:
  - a reversible binding with $g(T, R_C) = -\gamma RT + \kappa R_C$ and $h(T, R_C) = -g(T, R_C)$,
  - an irreversible binding following $g(T, R_C) = -\gamma RT$ and $h(T, R_C) = \gamma RT - \delta_{R_c} R_C$.

The five models in the collection resulting from the general model, their assumptions and number of parameters are given in Fig. 7.

The five models are calibrated to each of the four datasets. For the four datasets, the residual sum of squares between experimental and simulated data for model $i \in \{1, \ldots, 5\}$ is minimized using a genetic algorithm. For each dataset, the residual sum of squares is defined as

$$\mathrm{RSS}_i = \sum_{j=1}^{m} \left[\left(T_p^{exp}(t_j) - T_p^i(t_j) \right)^2 + \left(T_{total}^{exp}(t_j) - T_{total}^i(t_j) \right)^2 + \left(N_{total}^{exp}(t_j) - N_{total}^i(t_j) \right)^2 \right], \tag{8}$$

Model	p	rapamycin	feedback	mTOR
NTRr	12	reversible	no	s/d
NTtRr	11	reversible	no	constant
NTRi	12	irreversible	no	s/d
NTRrf	13	reversible	yes	s/d
NTRif	13	irreversible	yes	s/d

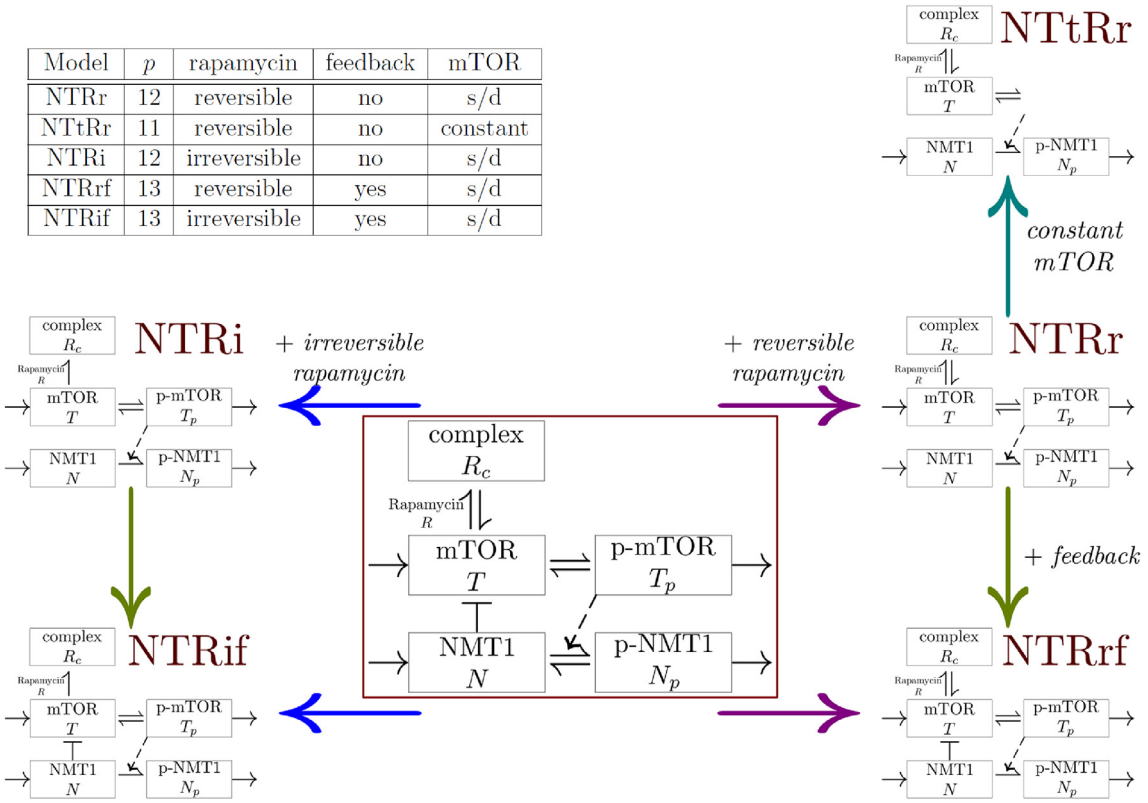

Fig. 7. General model and collection of models considered for the interactions between NMT1 and mTOR in breast cancer cells in Jacquier et al. (2018) (Jacquier et al., 2018). In the centre, the framed diagram is the general model. The non-framed diagrams represent the five models of the collection. At the top left corner, the table lists the collection of models studied with their number of parameters p, assumptions on rapamycin binding, feedback regulation of mTOR by NMT1 and mTOR dynamics (s/d meaning an explicit synthesis and degradation of mTOR). Figure adapted from (Jacquier et al., 2018).

where t_j are the $m = 9$ time points, $T^i_{total}(t_j) = T^i(t_j) + T^i_p(t_j)$ and $N^i_{total}(t_j) = N^i(t_j) + N^i_p(t_j)$ are the responses in model i. Hence, for a given dataset, a model is associated with a nominal set of parameters.

To identify the best model and characterize the required mechanisms for each of the four dataset, the corrected AIC are used as the number of observations is small ($n = 3m = 27$, for each dataset) in comparison to the number of parameters $p+1$ (see table in Fig. 7). For each dataset and each model $i \in \{1, \ldots, 5\}$, $AICc_i$ is computed as previously shown:

$$\mathrm{AICc}_i = n \ln\left(\frac{\mathrm{RSS}_i}{n}\right) + \frac{2k_i n}{n - k_i - 1}. \tag{9}$$

k_i is the number of estimated parameters for model i (p in the table displayed in Fig. 7), including the estimation of the variance by RSS_i/n. Using (7) with (9), the Akaike weights are computed for an easier interpretation of results (Table 1).

This example highlights that conclusions depend strongly on experimental data. For dataset 2 and 3, the Akaike weight of model NTtRr is found to be larger than 0.9; for dataset 2 and 3, model NTtRr is the best model. However, for dataset 1 and 4, none of models obtains a high Akaike weight. Model selection to find the best model is not conclusive when using dataset 1 and 4. However, discriminating between feedback and no feedback is conclusive for all the four datasets. There is strong evidence that there does not exist a negative feedback regulation of mTOR by NMT1. The reversible binding assumption is more likely to occur when dataset 2 and 3 are used; however, discriminating between reversible and irreversible binding is not conclusive when dataset 1 and 4 are considered.

In summary, the Akaike Information Criterion, valid to compare nested and non-nested models, can be used as a powerful tool for model selection. AIC or AICc allows the ranking of the candidate models and might select a best model within the collection given the experimental data considered. The best model has the lowest AIC or AICc values; however, the actual values of AIC or AICc have no meaning. The use of Akaike weights computed from AIC or AICc allows the interpretation of results by providing the probability that a model is the best model given the experimental data and the set of models considered. Furthermore, the Akaike weights allow the quantification of the likelihood of diverse assumptions considered within the model collection. The model selection method is specific to a given set of data; neither AIC nor AICc can be used to

Table 1
$AICc_i$ and Akaike weights w_i for the five models for datasets 1 to 4, with $k_i = p + 1$ the number of parameters considered to compute the $AICc_i$, with i denoting the model considered. The weights corresponding to each assumption are obtained by summing the weights of the models verifying the assumption (see the table in Figure 7). Thus, the weight of models with reversible (resp. irreversible) rapamycin effect is obtained by summing the weights of models including this assumption namely NTRr, NTtRr and NTRrf (resp. NTRi and NTRif). The weight of the feedback assumption corresponds to the sum of the weights of models NTRrf and NTRif.

Models		Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	k_i	$AICc_i$	w_i	$AICc_i$	w_i	$AICc_i$	w_i	$AICc_i$	w_i
NTRr	13	-2.3	0.393	-107.1	10^{-4}	-75.9	0.025	-58.8	0.407
NTtRr	12	-1.4	0.248	-124.9	**0.997**	-83.1	**0.929**	-58.2	0.305
NTRi	13	-2.0	0.351	-113.5	0.003	-76.7	0.039	-58.1	0.282
NTRrf	14	7.0	0.004	-91.6	10^{-8}	-73.2	0.007	-48.0	0.002
NTRif	14	6.6	0.005	-102.4	10^{-5}	-66.6	10^{-4}	-49.4	0.004
Assumptions									
Reversible			0.64		**0.9967**		**0.961**		0.714
Irreversible			0.36		0.0033		0.039		0.286
Feedback			0.008		10^{-5}		0.007		0.006
No feedback			**0.992**		**0.99999**		**0.993**		**0.994**

compare models based on different datasets. Information criteria are not statistical tests. Finally, never forget that the best model cannot be considered as the “truth”.

4. Conclusion

The principle of multiple working hypotheses is important in Science as it permits to confront different mechanisms; when considered with appropriate selection criteria, it allows the culling of hypotheses and the unimportant and required mechanisms can be identified. The use of a collection of models coupled with a model selection method instead of a single model makes the mathematical modelling approach more powerful and helps convince experimental scientists of the use and power of theoretical approaches. For instance, in (Kirmse et al., 2007), the use of a collection of models allows the identification of the longitudinal annealing of filaments as a required mechanism for the filament elongation of *in vitro* intermediate filaments. This mechanism was experimentally confirmed a few years later in (Colakoglu & Brown, 2009) and (Winheim et al., 2011). In (Portet, Madzvamuse, Chung, Leube, & Windoffer, 2015), the use of a collection of models combined with Akaike Information Criterion model selection allows to demonstrate that the directed transport of assembled intermediate filament proteins in cells was required to explain the spatial distributions of intermediate filaments.

Summing up, investigating a question in Biology by combining experimental data and mathematical modelling could follow the different steps (Baker et al., 2005; Burnham & Anderson, 2002; Miao et al.,2009 Symonds & Moussalli, 2011):

- a systematic modelling of possible scenarios based on biological hypotheses and first principles to design a collection of models;
- once the collection of models is formed, each model has to be calibrated using the same data set or data sets;
- once parameter estimates are known for each model, compute their AIC or AICc and associated Akaike weights; rank models and identify the best model or the 95% confidence set of models;
- partition the collection of models in subsets of models based on their underlying hypotheses and using Akaive weights, evaluate the importance of different processes.

Declaration of competing interest

No conflict of interest

Acknowledgements

SP is supported by a Discovery Grant of the Natural Sciences and Engineering Research Council of Canada (RGOIN-2018-04967).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Baker, C. T. H., Bocharov, G. A., Paul, C. A. H., & Rihan, F. A. (2005). Computational modelling with functional differential equations: Identification, selection, and sensitivity. *Applied Numerical Mathematics, 53*, 107–129.
Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.
Casement, R. (1984). *Man suddenly sees to the end of the universe*. University Press.
Coelho, F. C., Codeço, C. T., & Gomes, M. G. M. (2011). A Bayesian framework for parameter estimation in dynamical models. *PLoS One, 6*(5), 1–6.
Colakoglu, G., & Brown, A. (2009). Intermediate filaments exchange subunits along their length and elongate by end-to-end annealing. *Journal of Cell Biology, 185*, 769–777.
Friston, K. J. (2002). Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage, 16*, 513–530.
Gotoh, T., Kim, J. K., Liu, J., Vila-Caballer, M., Stauffer, P. E., Tyson, J. J., et al. (2016). Model-driven experimental approach reveals the complex regulatory distribution of p53 by the circadian factor period 2. In *Proceedings of the national academy of sciences*.
Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297.
Jacquier, M., Kuriakose, S., Bhardwaj, A., Zhang, Y., Shrivastav, A., Portet, S., et al. (2018). Investigation of novel regulation of n-myristoyltransferase by mammalian target of rapamycin in breast cancer cells. *Scientific Reports, 8*, 12969.
Janzén, D. L. I., Bergenholm, L., Jirstrand, M., Parkinson, J., Yates, J., Evans, N. D., et al. (2016). Parameter identifiability of fundamental pharmacodynamic models. *Frontiers in Physiology, 7*, 590.
Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution, 19*, 101.
Kirmse, R., Portet, S., Mücke, R., Aebi, U., Herrmann, H., & Langowski, J. (2007). A quantitative kinetic model for the in vitro assembly of intermediate filaments from tetrameric vimentin. *Journal of Biological Chemistry, 282*, 18563–18572.
Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79–86.
Miao, H., Dykes, C., Demeter, L. M., & Wu, H. (2009). Differential equation modeling of HIV viral fitness experiments: Model identification, model selection, and multimodel inference. *Biometrics, 65*, 292–300.
Pitt, J. A., & Banga, J. R. (2019). Parameter estimation in models of biological oscillators: An automated regularised estimation approach. *BMC Bioinformatics, 20*, 82.
Portet, S. (2015). Studying the cytoskeleton: Case of intermediate filaments. *Insights, 8*(2), 1–9.
Portet, S., Madzvamuse, A., Chung, A., Leube, R. E., & Windoffer, R. (2015). Keratin dynamics: Modeling the interplay between turnover and transport. *PLoS One, 10*(3), e0121090.
Sagar, A., LeCover, R., Shoemaker, C., & Varner, J. (2018). Dynamic optimization with particle swarms (DOPS): A meta-heuristic for parameter estimation in biochemical models. *BMC Systems Biology, 12*, 87.
Sugiura, N. (1978). Further analysis of the data by Akaike's Information Criterion and the finite corrections. *Communications in Statistics - Theory and Methods, 7*(13).
Symonds, M. R. E., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology, 65*, 13–21.
Winheim, S., Hieb, A. R., Silbermann, M., Surmann, E.-M., Wedig, T., Herrmann, H., et al. (2011). Deconstructing the late phase of vimentin assembly by total internal reflection fluorescence microscopy (TIRFM). *PLoS One, 6*, e19202.
Yates, J. W. (2006). Structural identifiability of physiologically based pharmacokinetic models. *Journal of Pharmacokinetics and Pharmacodynamics, 33*, 421–439.
Zhang, X., Cao, J., & J Carroll, R. (2015). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics, 71*, 131–138.