Beyond mathematical analysis

Stéphanie Portet

Department of Mathematics
University of Manitoba
Winnipeg, MB, Canada
Stephanie.Portet@umanitoba.ca

February 2024

Once the model is written

Mathematical analysis: identify the type of mathematical techniques and theories required for the analysis of the model.. and characterize the behavior of the model

Numerical experiment: conduct numerical simulations of the model..

Model calibration: identify and estimate the values of parameters..

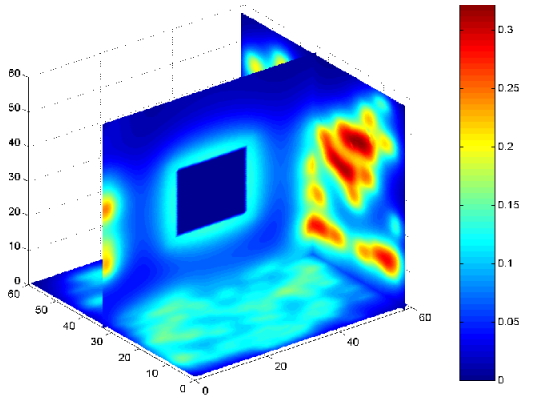
Sensitivity analysis: understand the effect of model inputs (parameters or initial conditions) on model outputs.. which parameter is the key driver of the model responses

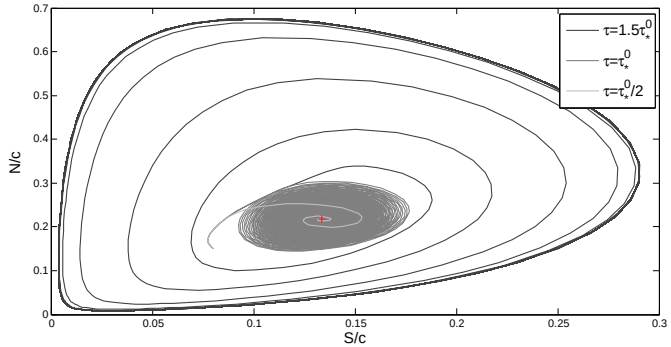
Validation: model must represent accurately the real process, it must reproduce known states of the real process.. if several models are considered, model selection has to be used.

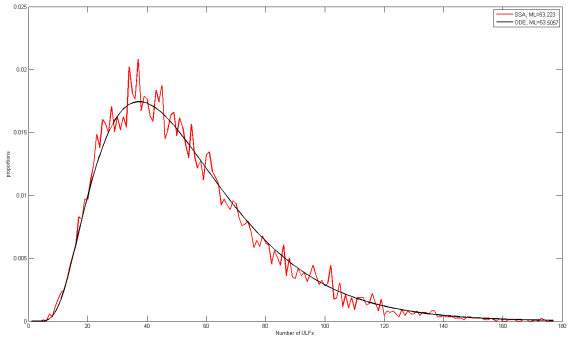
Outline

Numerical experiments to simulate solutions of models

- Approximation of solutions (Euler, Runge-Kutta, Euler-Maruyama, Milstein method..)
- Approximation of equations (Finite differences, finite elements method..)
- Computational methods
  - Agent-Based Models
  - Stochastic Simulation Algorithms (Gillespie, first reaction method, tau-leap..)







$\Rightarrow$ parameter values?

Gillespie's algorithm

Stochastic simulation algorithm (SSA) for constructing an exact numerical realization of the process $\mathbf{X}(t)$:

1. Initialize the time $t = t_0$ and the system's state $\mathbf{x} = \mathbf{x}_0$.
2. With the system in state $\mathbf{x}$ at time t, evaluate all the $a_j(\mathbf{x})$ and their sum $a_0(\mathbf{x}) = \sum_{j=1}^{M} a_j(\mathbf{x})$.
3. Generate 2 random numbers r_1 and r_2
4. Calculate the time to next event τ and the next event j using

$$\tau = \frac{-\ln(r_1)}{a_0(\mathbf{x})}$$

   and $j =$ the smallest integer satisfying $\sum_{j'=1}^{j} a_{j'}(\mathbf{x}) > r_2 a_0(\mathbf{x})$.
5. Effect the next reaction by replacing $t + \tau \to t$ and $\mathbf{x} + \mathbf{v}_j \to \mathbf{x}$.
6. Record $(\mathbf{x}, t)$ as desired. Return to Step 2, or else end the simulation.

A very simple chemical reaction: degradation

$$X \xrightarrow{k} \emptyset$$

Deterministic approach: $X(t)$ concentration of molecule X at time t

$$\frac{dX}{dt} = -kX, \quad X(0) = x_0, \Rightarrow X(t) = x_0 \exp^{-kt}.$$

Stochastic approach: $X(t)$ number of molecule X at time t (random variable), $p(x, t|x_0, t_0) := Prob\{X(t) = x, given\, X(t_0) = x_0\}$

$$\frac{dp(x, t|x_0, t_0)}{dt} = k(x+1)p(x+1, t|x_0, t_0) - k(x)p(x, t|x_0, t_0).$$

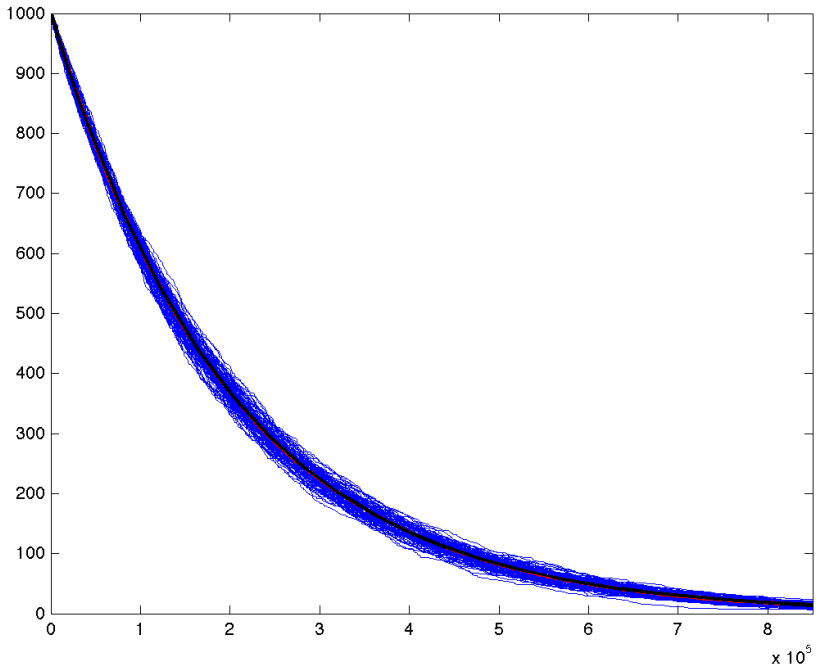
The solution is (Binomial probability density function)

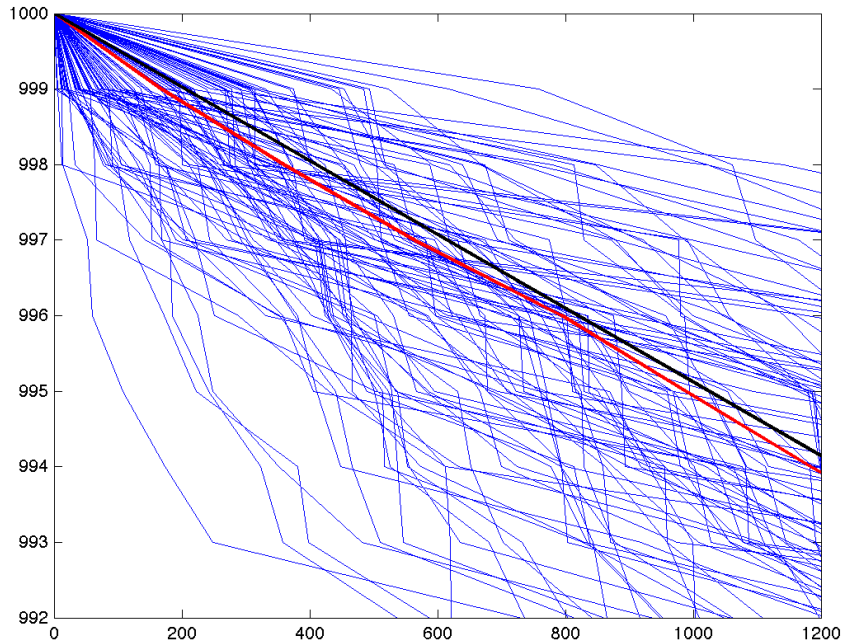
$$p(x, t|x_0, 0) = \frac{x_0!}{x!(x_0 - x)!} \exp^{-kxt}(1 - \exp^{-kt})^{x_0 - x}, \quad (x = 0, \ldots, x_0)$$

$$E(X) = x_0 \exp^{-kt}, \quad Var(X) = x_0 \exp^{-kt}(1 - \exp^{-kt})$$

$$X \xrightarrow{k} \emptyset$$

Simulations done with Gillespie's algorithm

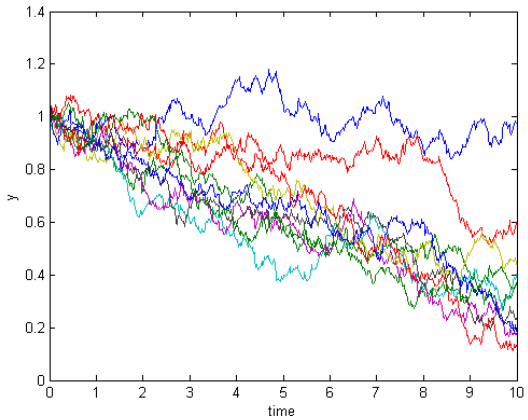


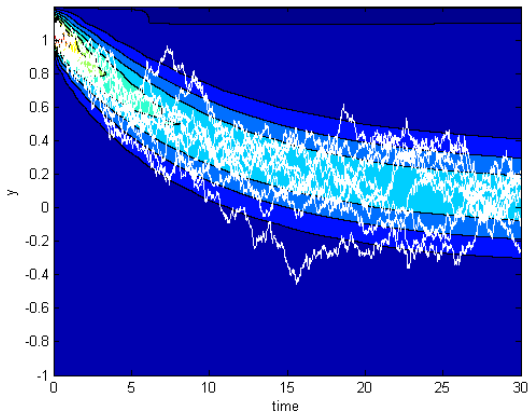


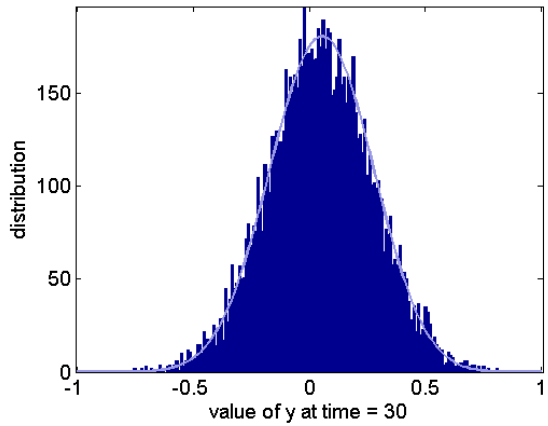
blue = realization, red = mean of realizations, black = ODE solution

Individual vs Population

Langevin equation: (stochastic equation of motion for the time evolution of y) motion of a particle in a viscous medium subject to friction and noise (white Gaussian noise).
Fokker-Planck equation: equation of motion for the time dependent probability distribution $p(y, t)$.







Parameters

To run numerical experiments, parameter values need to be chosen/determined

- known from the literature
- measured experimentally
- unknown.. $\Rightarrow$ Model Calibration

Outline

Model calibration/Estimation of parameters

Consider a model

$$\frac{dx}{dt} = m(t, x, p)$$

- t: independent variable
- x: vector of state variables
- p: vector of parameters

Parameters of the model

- have (biological) interpretations
- and their values are unknown

$\Rightarrow$ necessary to estimate (find appropriate values) the model parameters from measurements in presence of errors in measurements

Method: Fitting the solutions of ODEs to data

Model calibration/Estimation of parameters/Data fitting

Two general (optimal) methods of parameter estimation based on

- Least Squares (LS)
  - minimizing the objective function = sum of squared residuals of all measurements
- Maximum Likelihood (ML)
  - maximize the likelihood function

If the measurement errors are independent, normally distributed with a common variance, the LS and ML methods are equivalent (give the same result)

Optimization problems

When an analytic expression of the function $\Phi(p)$ to optimize is known

Theorem

A smooth function $\Phi(p)$ attains an local minimum (resp. maximum) at $\hat{p}$ if

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

$$z^T H(p) z > 0 (\text{ resp. } < 0)$$

where z is any real vector.

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

Outline

Curve fitting: method of least-squares

Model: $f(x, p)$ with parameters p
Criterion: measure the total error in fitting a curve to data

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

- sum of squares for error = sum of squared residuals
- residual = difference between the actual value (data) and the predicted value (curve)

Aim: minimization of the sum of squared residual

$$\min_p RSS(p)$$

Result: Least-squares best fit minimizes the sum of squares of vertical distances between data points and fitting curve points.

Method of least-squares for models linear in parameters

Aim: find parameter values for the model which best fits data

- **Observation**: n data points (x_i, y_i) with $i = 1, \cdots, n$
- **Model**: $f(x, p)$ where p is a $m-$vector of parameters (m parameters)
- **Criterion**: sum, RSS, of squared residuals

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

- **Solution**: $\hat{p}$ such that

$$RSS(\hat{p}) = \min_{p} RSS(p)$$

is obtained by setting the gradient equal to zero (m parameters)

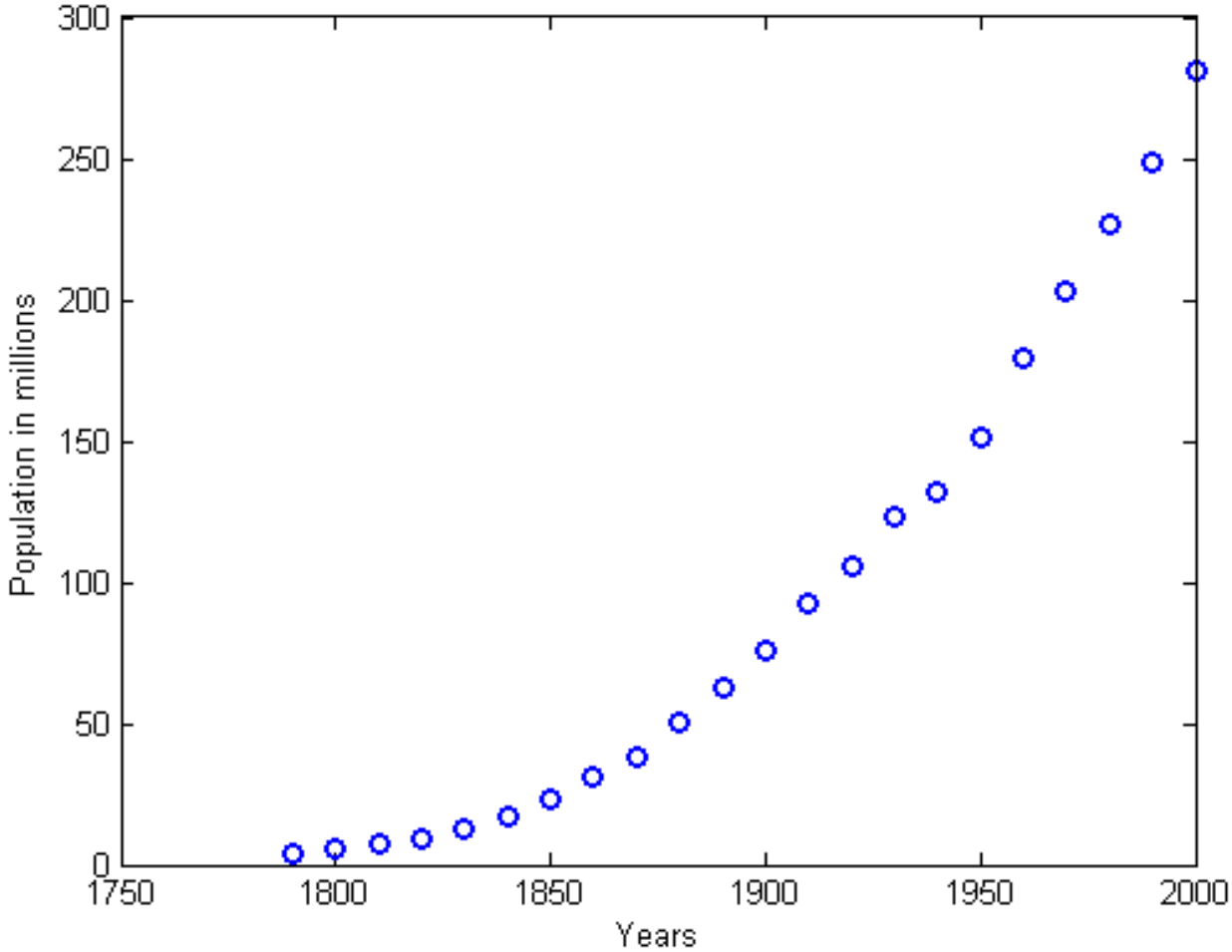
$$\frac{\partial RSS}{\partial p_j} = 0, \quad j = 1, \cdots, m$$

or

$$-2 \sum_{i=1}^{n} (y_i - f(x_i, p)) \frac{\partial f(x_i, p)}{\partial p_j} = 0, \quad j = 1, \cdots, m$$

US population (in millions) from 1790 to 2000

1790	3.929
1800	5.308
1810	7.240
1820	9.638
	12.866
	17.069
	23.192
	31.443
	38.558
	50.156
	62.948
1900	76.212
	92.228
	106.021
	123.202
	132.164
	151.325
	179.323
	203.302
	226.542
	248.709
2000	281.421



Observation: 22 data points $(x_i, y_i) =$ (year,population) with $i = 1, \ldots, 22$

US population (in millions) from 1790 to 2000

Model: $f(x, p)$ where p is a $k-$vector of parameters (k parameters)

Hypothesize the form of the function f

- **Quadratic function** (x years)

$$f(x) = y = a + bx + cx^2$$

$k = 3$ parameters to estimate a, b and c

- **Exponential function** (x years)

$$f(x) = y = a \exp^{bx}$$

Change of variable $\ln y = Y$

$$\ln y = Y = \ln a + bx = A + bx$$

$k = 2$ parameters to estimate A and b

Both models are linear in parameters

US population (in millions) from 1790 to 2000

Find the minimum of

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

Set the gradient of RSS to zero

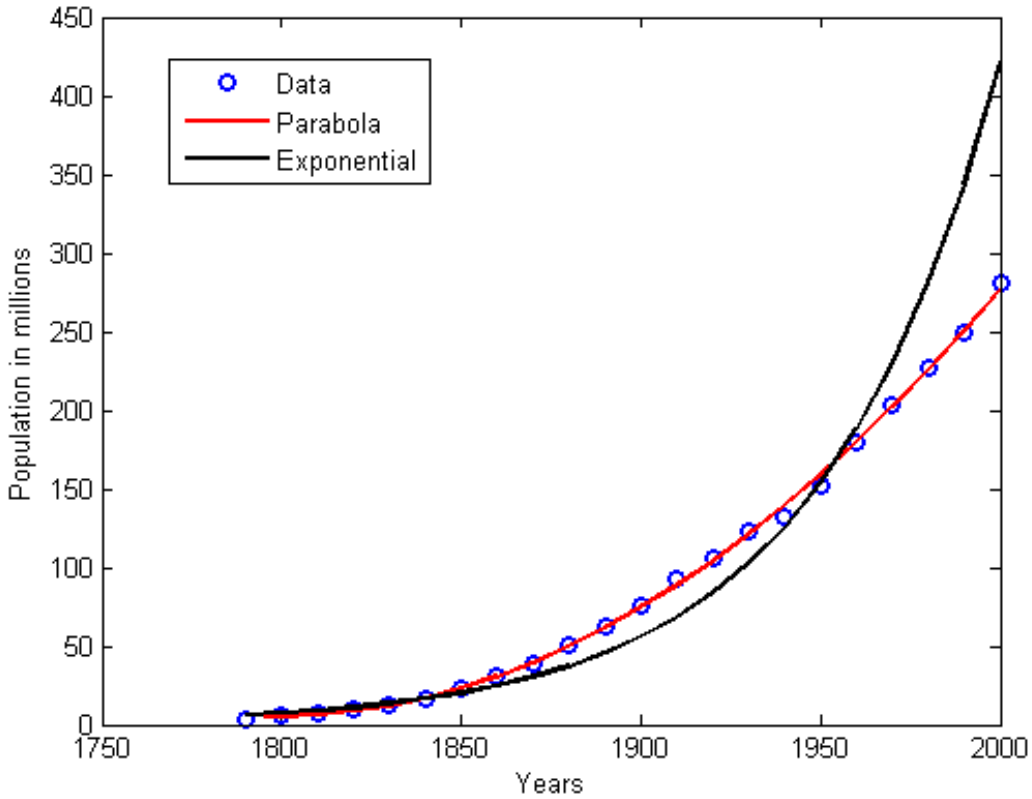
$$\sum_{i=1}^{n} (\ln y_i - (A + bx_i)) \frac{\partial (A + bx_i)}{\partial A} = \sum_{i=1}^{n} (\ln y_i - (A + bx_i)) = 0$$

$$\sum_{i=1}^{n} (\ln y_i - (A + bx_i)) \frac{\partial (A + bx_i)}{\partial b} = \sum_{i=1}^{n} (\ln y_i - (A + bx_i)) x_i = 0$$

$\hat{A}$ and $\hat{b}$ (estimate of A and b) satisfy

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \hat{A} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \ln y_i \\ \sum_{i=1}^{n} x_i \ln y_i \end{bmatrix}$$

US population (in millions) from 1790 to 2000



Quadratic curve (x years)

$$f(x) = 0.0067x^2 - 24.0358x + 21620.47$$

Exponential curve (x years)

$$f(x) = 1.2162 \times 10^{-15} exp^{0.0202x}$$

Naive approach to compare models: R^2

Measure of the goodness of fit

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

where

- RSS = residual sum of squares
- n = sample size
- y = data

Select the model that maximizes R^2

Best fit but neglect the model complexity (select the more parameter rich model)

Only valid for linear models

To compare models: Ajusted R^2

Replacing the two variances with their unbiased estimates

Measure of the goodness of fit

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{\sum(y_i - \bar{y})^2/(n-1)}$$

where

- RSS = residual sum of squares
- n = sample size
- y = data
- p = number of parameters

Select the model that maximizes R^2_{adj}

Only valid for linear models

Estimation of parameters in mechanistic models

$$\frac{dx}{dt} = m(x, p, t),\ x(t_0) = x_0(p),\ \tilde{y} = h(x, p, t)$$

$x(t)$ vector of state variables, x_0 IC, h observable function and p vector of unknown constant parameters

To find the vector of parameter values p that minimizes the distance between measured observations and simulated observations:

- Define a "distance" = Scalar objective function (cost function)

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

  n_e # of experiments, n_o^e # of observable per experiments, $n_i^{e,o}$ # of samples per observable per experiments
  $y_e^o(t_i)$ measured data, $\omega_i^{e,o}$ weights and $\tilde{y}_o^e(t_i, p)$ simulated output

- Optimization method to minimize $F_{ls}(p)$ to find $\hat{p}_{LSE}$

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

Optimization methods

When an analytic expression of the function to optimize is unknown

- Local optimization methods:
  - gradient descent-based methods: Levenberg-Marquardt or Gauss-Newton
  - derivative-free local search methods: Nelder-Mead method
  - only find a global optimum for appropriate starting points
  - converge to local optima
  - suboptimal solutions
- Global optimization methods:
  - simulated annealing
  - genetic algorithm
  - particle swarm

Pitt and Banga (2019) BMC Bioinformatics. 20:82. Sagar *et al.* (2018) BMC Systems Biology 12:87.

Outline

Maximum Likelihood Estimator Approach

- Experimental data $y = (y_1, \ldots, y_n)$ = random sample generated from an unknown probability distribution function (pdf) depending on parameters $p = (p_1, \ldots, p_k)$
- Model = 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

$$\begin{aligned} f((y_1, \ldots, y_n)|(p_1, \ldots, p_k)) = & f(y_1|(p_1, \ldots, p_k)) f(y_2|(p_1, \ldots, p_k)) \\ & \cdots f(y_n|(p_1, \ldots, p_k)) \end{aligned}$$

(function of y = the probability of data y for a given value of p)
Varying the parameter p across its range defines a model (a family of pdfs)

Likelihood function

Problem: Find the pdf among the all pdfs of the family that is the most "likely" to have produced the data (inverse problem)

The likelihood function $\mathcal{L}$ is the density function regarded as a function of p

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

- The likelihood of a particular value of 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.
- The density function f gives the probability of observing y given the parameter p and sums to 1 over all the possible values of y. (function of y, data scale)
- The likelihood function $\mathcal{L}$ is a function of p given the data and does NOT sum to 1 over the possible values of p. (parameter scale)

Maximum Likelihood Estimation (MLE)

Problem: Seek the value $\hat{p}_{MLE} = (\hat{p}_{1,MLE}, \ldots, \hat{p}_{k,MLE})$ of the parameter vector p that maximizes the likelihood function $\mathcal{L}(p|y) \Leftrightarrow$ Maximize the log-likelihood $\ln \mathcal{L}(p|y)$ to find Maximum Likelihood Estimator $\hat{p}_{MLE}$

$\hat{p}_{MLE}$ satisfy the following conditions:

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

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

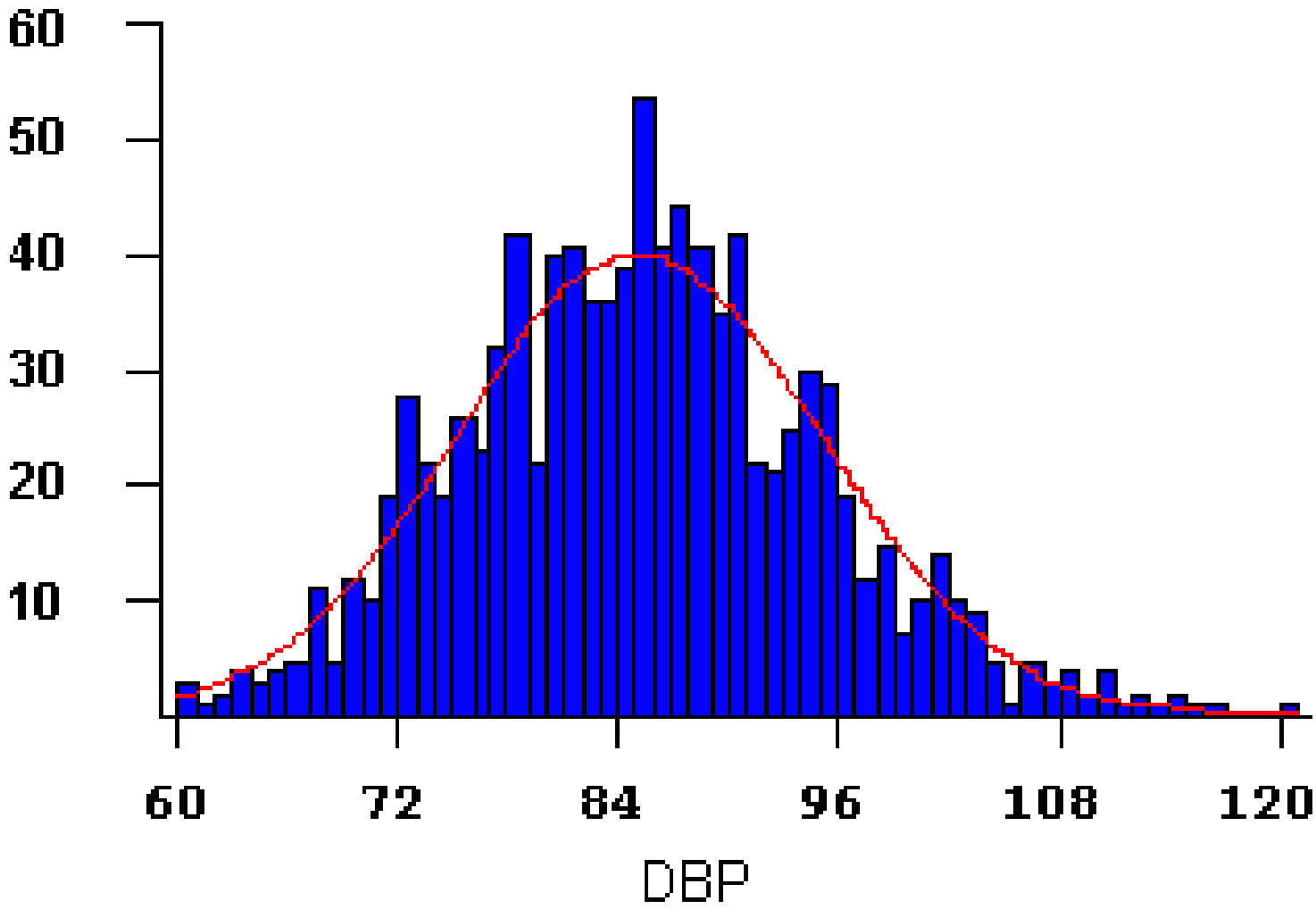
- Convexity condition: consider the Hessian matrix $H(p)$, $H_{i,j}(p) = \frac{\partial^2 \ln \mathcal{L}(p|y)}{\partial p_i \partial p_j}$ with $i, j \in \{1, \ldots, k\}$.

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

where z is any $k \times 1$ real vector.

Example: Normally distributed data

Data: Blood pressure of 1000 patients



Assumption: Data are normally distributed

Example: Normally distributed data

Assume that experimental data $y_1, \ldots, y_n$ is drawn from a $\mathcal{N}(\mu, \sigma^2)$ with μ and σ unknown ($k = 2$, $p = (\mu, \sigma)$).
Let $Y_1, \ldots, Y_n$ be n i.i.d.[1] $\mathcal{N}(\mu, \sigma^2)$ random variables,

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

the joint pdf is

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

For a fixed data set $y_1, \ldots, y_n$, the likelihood function with $\mathcal{L}(\mu, \sigma|y_1, \ldots, y_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(y_i-\mu)^2}{2\sigma^2}}$ and the log-likelihood function

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

[1] independent and identically distributed

To find $\hat{\mu}_{MLE}$

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

To find $\hat{\sigma}_{MLE}$

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

using $\mu = \hat{\mu}_{MLE}$ we have $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{MLE})^2$

Maximum Likelihood Estimates are $\hat{\mu}_{MLE}(= 85)$ the mean of data and $\hat{\sigma}^2_{MLE}(= 20^2)$ the variance of data and

$$\ln \mathcal{L}(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}|y_1, \ldots, y_n) = \max_{\mu,\sigma} \ln \mathcal{L}(\mu, \sigma|y_1, \ldots, y_n)$$

Outline

Estimation of parameters in mechanistic models (LS)

To find the vector of parameter values p that minimizes the distance between measured observations and simulated observations:

- $y_e^o(t_i)$ measured data (n_e # of experiments, n_o^e # of observable per experiments, $n_i^{e,o}$ # of samples per observable per experiments)
- $\frac{dx}{dt} = m(x, p, t),\ x(t_0) = x_0(p),\ \tilde{y} = h(x, p, t)$
  h observable function and p vector of unknown constant parameters
  $\Rightarrow \tilde{y}_o^e(t_i, p)$ simulated output
- "Distance" = Scalar objective function (cost function)

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

  $\omega_i^{e,o}$ weights
- Optimization method to minimize $F_{ls}(p)$ to find $\hat{p}_{LSE}$

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

Relationship between LSE and MLE

Assuming **independent, normally distributed additive measurement error** (noise) with standard deviation $\sigma_i^{e,o}$, the probability of observing the data y given the parameters p and $\sigma_i^{e,o}$ is

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

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 $\sigma_i^{e,o}$. Maximizing the likelihood is equivalent to minimize the negative of the log-likelihood function:

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

Relationship between the negative log-likelihood and least-squares cost function when $\sigma_i^{e,o} = \sigma$ and $\omega_i^{e,o} = 1$ (ordinary LS):

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

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

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

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

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

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

Confidence intervals (CIs) for parameters

- (Wilks' theorem) The opposite of twice the logarithm of the likelihood ratio statistic (N_{param} : # of parameters)

$$-2\ln\left(\frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta}_{MLE})}\right) \rightsquigarrow \chi^2_{N_{param}}$$

- The two-sided $100(1-\alpha)\%$ confidence interval for θ consists of all θ values associated with the log-likelihood function being at a distance of less than $\chi^2_{N_{param},\alpha}/2$ of its peak value at the maximum likelihood estimator (N = number of observations)

$$\ln\left(\mathcal{L}(\theta)\right) \geq \ln\left(\mathcal{L}(\hat{\theta}_{MLE})\right) - \frac{\chi^2_{N_{param},\alpha}}{2}$$

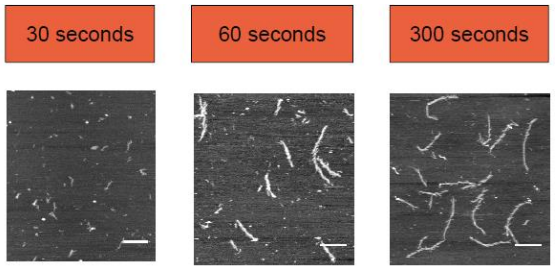
$$CI = \left\{\theta : \ln\left(\mathcal{L}(\hat{\theta}_{MLE})\right) - \ln\left(\mathcal{L}(\theta)\right) \leq \frac{\chi^2_{N_{param},\alpha}}{2}\right\}$$

Under the previous assumption:

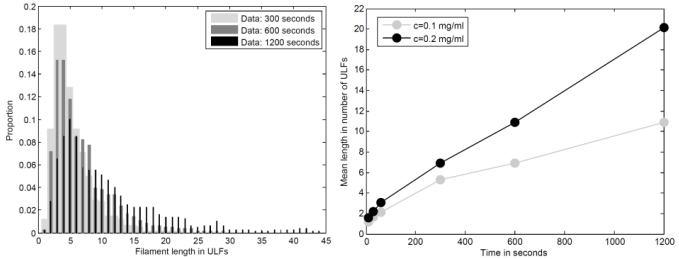
$$CI = \left\{\theta : \ln\left(\frac{F_{ls}(\theta)}{N}\right) - \ln\left(\frac{F_{ls}(\hat{\theta}_{LSE})}{N}\right) \leq \frac{\chi^2_{N_{param},\alpha}}{N}\right\}$$

Example: Dynamics of *in vitro* assembly of proteins

Experimental data



- In vitro assembly of filaments
- Imaging: EM, SFM, AFM, TIRF

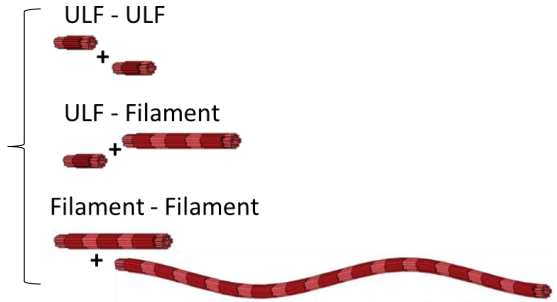


- Filament length distributions over time
- Mean lengths over time
- Lengths in # of ULFs

Model

- ULF = smallest structural stage
- Longitudinal annealing only

$$F_i + F_j \overset{k_0 \times k_{i,j}}{\longrightarrow} F_{i+j}$$



Effective association rates = k_0 x $k_{i,j}$

- $k_{i,j}$ = diffusion-controlled association rates
- k_0 = intrinsic bimolecular rate constant (single free parameter)

Smoluchowski's coagulation equation =>

- F_i concentration of filaments composed of i ULFs over time
- Filament length distributions over time
- Mean lengths over time
- Lengths in # of ULFs

Find k_0 x $k_{i,j}$ to best represent the data

Work in collaboration with H. Herrmann, N. Mücke (DKFZ, Heidelberg)

Example: Dynamics of *in vitro* assembly of proteins

Model depends on a single unknown parameter k_0

$$\frac{1}{k_0}\frac{\mathrm{d}F_i}{\mathrm{d}t} = \frac{1}{2}\sum_{j=1}^{i-1}(1+\delta_{j,i-j})\,k_{j,i-j}\;F_jF_{i-j} - \sum_{j=1}^{2N}(1+\delta_{j,i})\,k_{j,i}\;F_jF_i$$

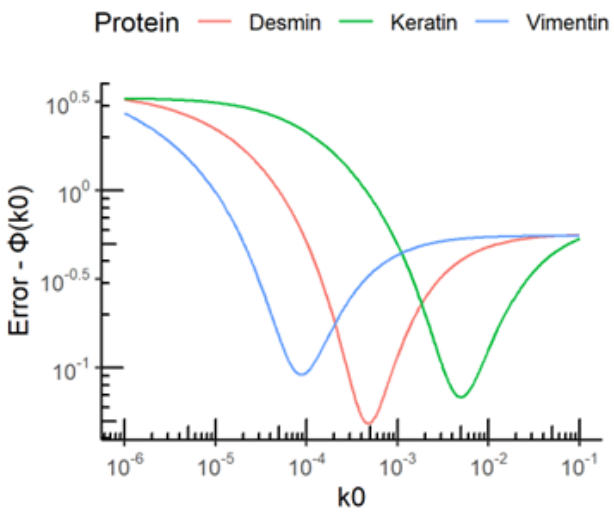
Observable function

$$ML(t) = \sum_{i=1}^{N} i\left[\frac{F_i(t)}{\sum_{i=1}^{N}F_i(t)}\right]$$

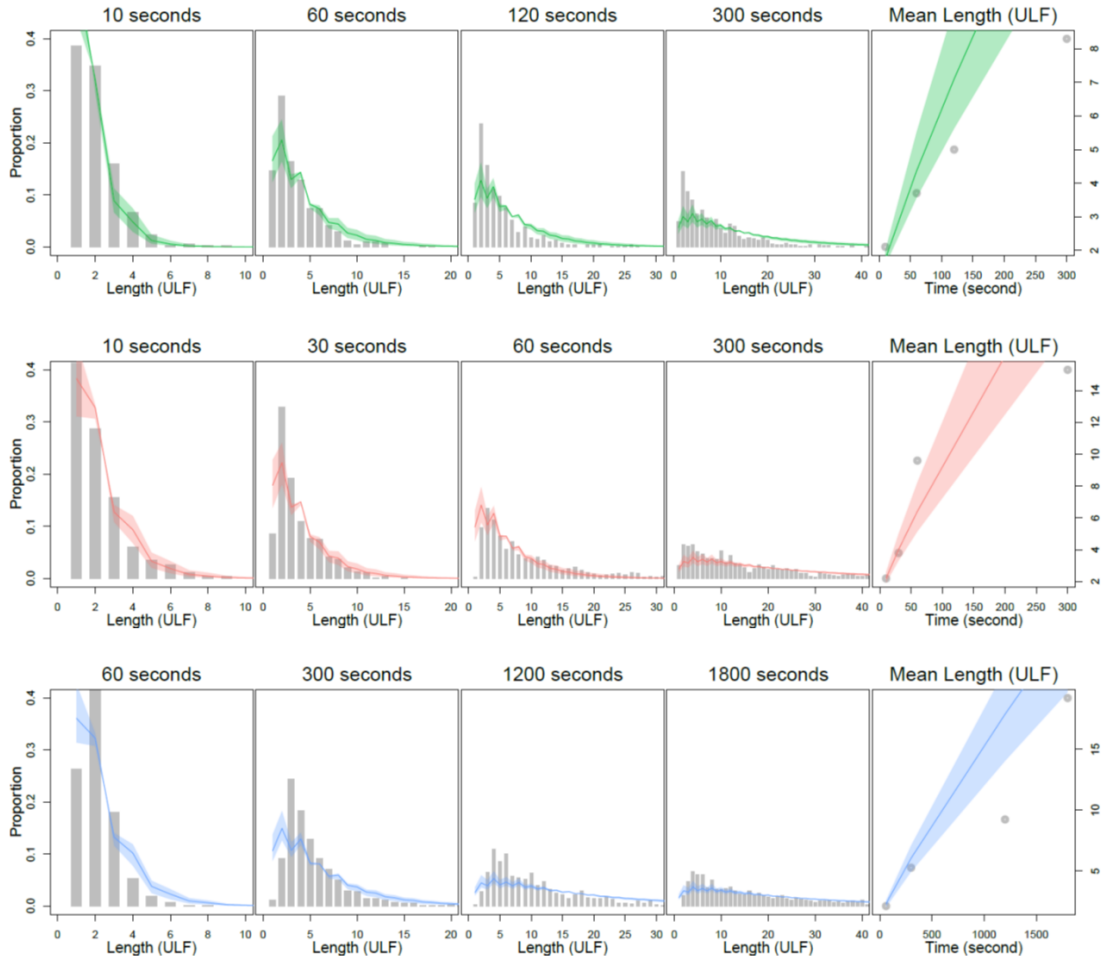
Error

$$\Phi(k_0) = \sum_{j=1}^{M}(ML_{model}(t_j,p) - ML_{data}(t_j))^2$$

Example: Dynamics of *in vitro* assembly of proteins



	Keratin	Desmin	Vimentin
90% IC	$[3.6\text{x}10^{-3}, 7\text{x}10^{-3}]$	$[3.5\text{x}10^{-4}, 6.7\text{x}10^{-4}]$	$[6.6\text{x}10^{-5}, 1.1\text{x}10^{-4}]$
$\widehat{k_0}$	$5\text{x}10^{-3}$	$4.8\text{x}10^{-4}$	$8.9\text{x}10^{-5}$



Model selection: Akaike Information Criterion

In 1973, Akaike found a relationship between the maximum likelihood (statistical analysis) and Kullback-Leibler divergence (to measure the difference between two probability distributions (information theory))

Akaike Information Criterion

$$AIC = -2 \ln \left(\mathcal{L}(\hat{\theta}_{MLE}|y) \right) + 2K$$

where $\mathcal{L}$ is the likelihood function (proba of obtaining the observed data y if the parameter had a given value), $\hat{\theta}_{MLE}$ is the maximum likelihood estimate of θ and K is the number of estimated parameters

Best model = the one with minimum AIC value

Akaike (1973) In: Petrov BN, Csaki F (eds) Second international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281

When the **measurement errors are independent and identically normally distributed with the same variance**

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

where $\hat{p}_{MLE} = \hat{p}_{LSE}$ $(N = n_e n_o^e n_i^{e,o})$

When **the data used to compare all the models are the same**, AIC can be computed as follows

$$AIC = N\ln\left(\frac{F_{ls}(\hat{p}_{LSE})}{N}\right) + 2K$$

where K is the number of estimated parameters (number of estimated mathematical model parameters + 1), N is the number of observations

Compute AIC_i of each model i with $i \in \{1, \ldots, R\}$
Best model = the one with minimum AIC value

AIC differences

As only the estimates of the expected relative K-L 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.

AIC differences: estimate of information loss when using model i rather than the estimated best model

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

$\min_i AIC_i$ = AIC of the best model in the collection

Interpretation = the larger the Δ_i, the less plausible is model i

Akaike (1974) IEEE Trans. Automatic Control. 19:716. Burnham and Anderson (2002) Model selection and multimodel inference: a practical information-theoretic approach. Second Edition. Springer.

Interpretation = the larger the Δ_i, the less plausible is model i

Some guidelines for nested models:

- $\Delta_i \in \{1, 2\}$ 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, can be omitted

Might be different for non-nested models or for a very large number of models

Burnham and Anderson (2002) Model selection and multimodel inference: a practical information-theoretic approach. Second Edition. Springer.

Akaike weights

For an easier interpretation, rescaling of Δ_i

Likelihood of model i given the data $\propto \exp\left(-\frac{\Delta_i}{2}\right)$

Akaike weight or "weight of evidence" of model i for being the best model of the collection given the data

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

w_i = probability that model i is the best (approximating) model given the experimental data and the collection of models considered

Interpretation

- The smaller the weight w_i, the less plausible is model i
- Consider a single best model i if $w_i > 0.9$

Uses of Akaike weights

Evidence ratio of model i versus model j = Strength of evidence in favour of model i over model j

$$\frac{w_i}{w_j}$$

Confidence set of models:
Sum the Akaike weights from largest to smallest until the sum is ≥ 0.95 $\Rightarrow$ the corresponding subset of models is the 95% confidence set on the best model

Relative importance of a process:
Sum the Akaike weights over all models in which the process of interest appears = measure of the relative importance of the process of interest

Model selection with Akaike Information Criterion

- Best model = the one with the lowest AIC
- Best model within the collection of model considered given the experimental data $\neq$ "true model"
- No meaning in the actual values of AIC
- Ranking of candidate models
- Selection of a model with the least number of parameters that best-fits experimental data
- Specific to a given set of data (cannot be used to compare models on different data sets)
- Valid to compare nested or non-nested models
- Not a test!!

Portet (2020) Inf. Dis. Model. 5.

Strategy as using data and mathematical modelling

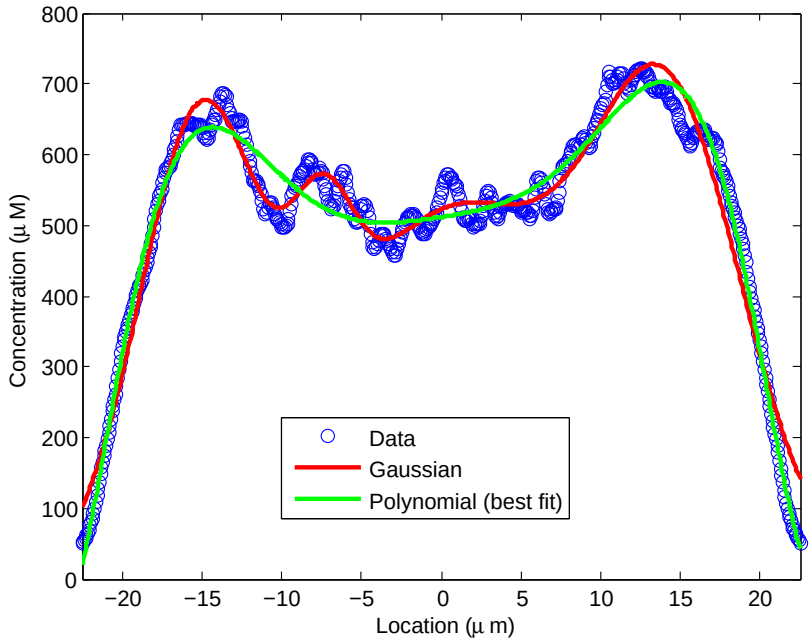
- Systematic modelling of possible scenarios based on biological hypotheses and first principles to design a collection of models
- Calibration of each model using **the same data**
- Compute AIC and Akaike weights for each model $\Rightarrow$ rank models and identify the best model or the 95% confidence set of models
- Partition the collection of models in subsets of models based on their underlying hypotheses and using Akaike weights, evaluate the importance of different processes

Portet *et al.* (2015) PLOS One. Jacquier *et al.* (2018) Scientific Report. Lee *et al.* (2021) AIMS Mathematics. Portet (2020) Infectious Disease Modelling.

Outline

Problems in parameter estimation

- lack of prior knowledge about parameters
- lack of identifiability
- convergence to local optima (ill-conditioning and non-convexity)
- overfitting (fitting the noise instead of signal)



Still an accurate fit is the starting point for uncertainty analysis or model selection..

Identifiability

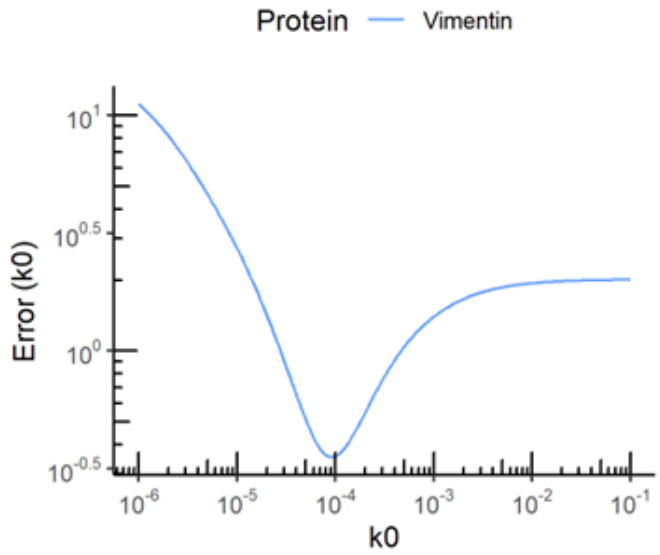
Can unknown model parameters uniquely be determined by parameter estimation from measured data? $\Rightarrow$ Identifiability

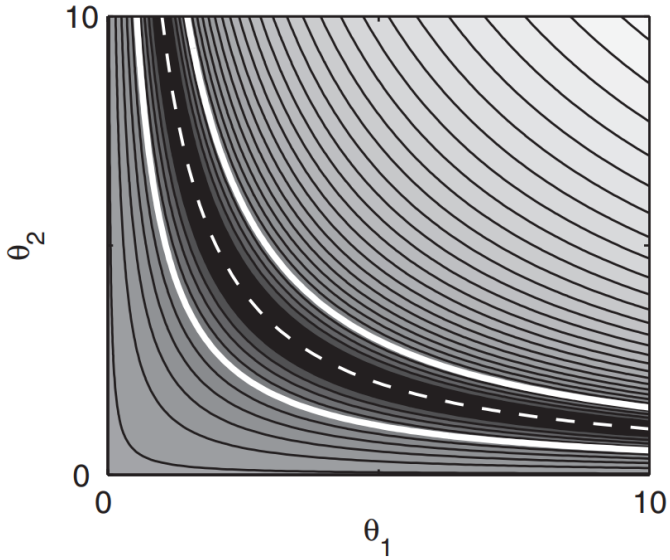
Two problems:

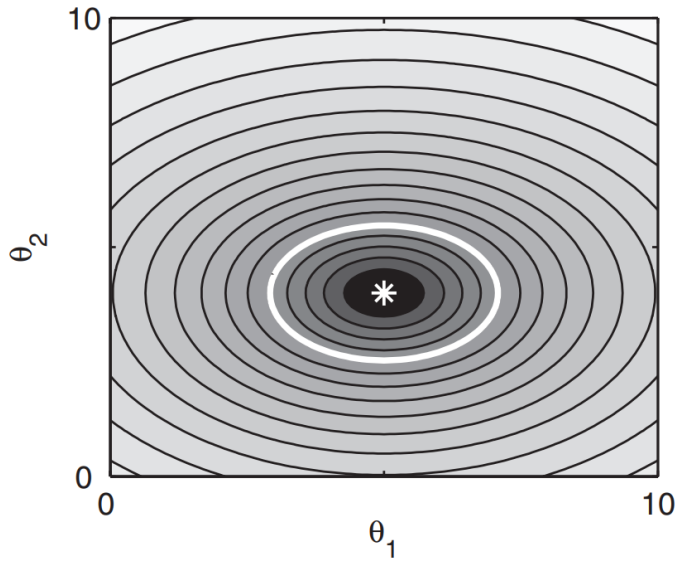
1. the larger the number of unknown parameters in a model, the larger the amount of quantitative data necessary to determine meaningful values for these parameters (Pratical identifiability)
2. even if appropriate experimental data are available, model parameters may not be uniquely identifiable (Structural identifiability)

Identifiability

Profiles of error as a function of parameters to be estimated







Outline

Sensitivity analysis

"Sensitivity analysis is the study of how the variation in the output of a model (numerical or otherwise) can be apportioned, qualitatively or quantitatively, to different sources of variation, and how the given model depends upon the information fed into it."

What are the effects of a parameter on the equilibrium value, an aggregated parameter, eigenvalues?

$\Rightarrow$ **Sensitivity analysis**

Sensitivity analysis is an important tool in studies of the dependence of systems on parameters and to identify the key drivers of the dynamics

Saltelli, A. *et al.* Sensitivity analysis. Wiley (2000). Saltelli, A. *et al.* Global sensitivity analysis: the primer. (2008)

Uncertainty analysis: Latin Hypercube Sampling (LHS)

A method of sampling that allows an unbiased estimate of the average model output, with the advantage that it requires fewer samples than simple random sampling to achieve the same accuracy.

LHS ensures that the entire parameter space is represented, that is, that the range of each parameter is fully stratified

Stratified sampling without replacement technique

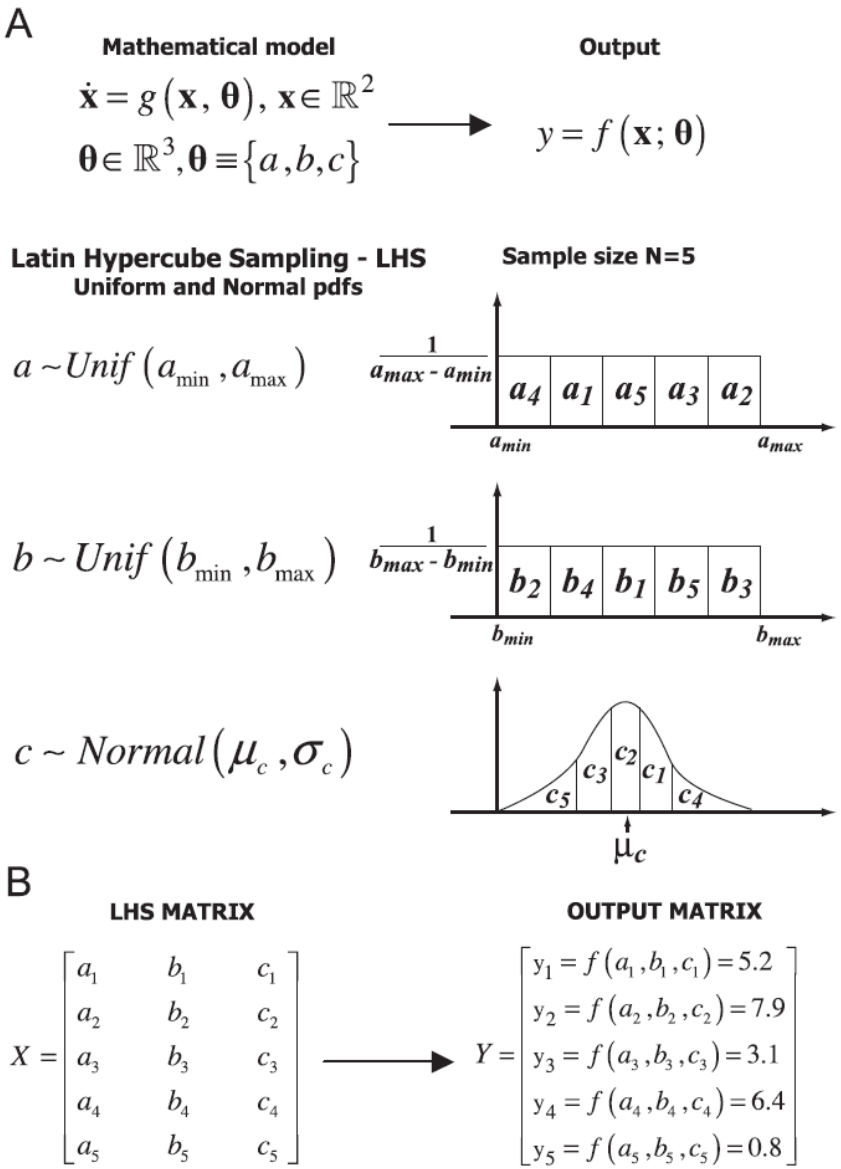
Parameters of LHS

- Specify a probability density function (pdf) for each parameter
- N = sample size (N is at least $k + 1$ where k is the number of parameters varied)
- Choose the model output(s)

Marino *et al.* J. Theor. Biol. (2008) 254:178. Renardy *et al.* Cur. Op. Biom. Eng. (2019) 11:109.

Latin Hypercube Sampling (LHS)

- Divide the random parameter distributions into N equal probability intervals, which are then sampled
- Sampling performed independently for each parameter
- Sampling is done by randomly selecting values from each pdf
- Each interval for each parameter is sampled exactly once



Marino *et al.* J. Theor. Biol. (2008) 254:178.

Sensitivity analysis

Local sensitivity analysis method: performing the analysis around a point of interest in the model input space (nominal parameter set) (derivative-based method)

Global sensitivity analysis methods: assuming probabilistic distributions for models inputs

- **Correlation-based methods**: evaluate effects of a factor p_i while the others are also allowed to vary. Partial rank correlation coefficient (PRCC) provides a measure of monotonicity after the removal of the linear effects of all but one variable.
- **Variance-based methods**: quantify the amount of variance that each input factor p_i contributes with on the unconditional variance of the output $V(y)$. Extended Fourier amplitude sensitivity test (eFAST) return measures of fractional variance accounted for by individual parameters and groups of parameters.

Marino *et al.* (2008) J. Theor. Biol. 254, 178. `http://malthus.micro.med.umich.edu/lab/usadata/`

Example of local sensitivity analysis: Discrete structured models

$$x(t+1) = Px(t)$$

- λ_1 dominant eigenvalue of P
- **Stable stage distribution** V_1 (right eigenvector associated to λ_1)

$$PV_1 = \lambda_1 V_1.$$

For P, any initial population stage structure projected forward will approach the stable stage distribution V_1, where each stage class increases in size λ_1 times each time period.

- **Reproductive value of each stage** W_1 (left eigenvector associated to λ_1)

$$W_1^T P = \lambda_1 W_1^T,$$

These reproductive values estimate the expected reproductive contribution of each stage to population growth.

- Total population $p(t) = \sum_{i=1}^{n} x_i(t)$

$$p(t) = \lambda_1^t p_0$$

How sensitive the population growth rate is to variations in fecundity, growth or survival rates?

$$x(t+1) = Px(t)$$

Sensitivity of λ_1 to changes in life cycle parameters

- proportional sensitivity (or "elasticity") of λ_1: proportional change in λ_1 caused by proportional change in one of the life cycle parameters.

$$\frac{P_{ij}}{\lambda_1}\frac{\partial \lambda_1}{\partial P_{ij}} = \frac{P_{ij}}{\lambda_1}\left(\frac{W_{1_i} V_{1_j}}{< W_1, V_1 >}\right)$$

Because these elasticities sum to 1, the relative contribution of the matrix elements to λ_1 can be compared.

Caswell (1982) Ecology 63(5) - 1223-1231.

Chinook salmon in the Columbia river

4 developmental stages, $x \in \mathbb{R}^4$:

- eggs
- yearlings
- 2-year olds
- 3-year olds

$$x(t+1) = Px(t), \quad P = \begin{bmatrix} 0 & 4 & 20 & 60 \\ 0.05 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \end{bmatrix}$$

The elasticities matrix is

$$\begin{bmatrix} 0 & 0.059 & 0.087 & 0.155 \\ 0.301 & 0 & 0 & 0 \\ 0 & 0.242 & 0 & 0 \\ 0 & 0 & 0.155 & 0 \end{bmatrix}$$

Derivative-based method - Forward sensitivity equations

$$\frac{dx}{dt} = f(t, x, p)$$

- x: vector of state variables
- p: vector of parameters

Sensitivity of variables to parameter p_i, $\frac{\partial x}{\partial p_i}$, satisfies

$$\frac{d}{dt}\frac{\partial x}{\partial p_i} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial p_i} + \frac{\partial f}{\partial p_i}$$

(forward sensitivity equations $\Leftarrow$ differentiation of original system with respect to p_i and inversion of differentiation operators)

Sensitivity coefficients

$$\frac{\partial x}{\partial p_i}$$

Ingalls *et al* (2003) J. Theor. Biol. 222, 23.

Forward sensitivity equations

Sensitivity coefficients represent

- derivatives of the model responses with respect to parameters
- rates of change of variables with respect to an increase in a given parameter p_i

To allow comparisons, normalized sensitivity coefficients are used:

$$\frac{p_i}{x}\frac{\partial x}{\partial p_i}$$

Possible approaches:

- Local sensitivity analysis: values of normalized sensitivity coefficients when parameters are set to their nominal values
- Global sensitivity analysis: create input factor distributions through sampling (e.g. Latin Hypercube Sampling)

Variance-based methods

First-order index represents main effect contribution of each input factor to the variance of the output

$$S_i = \frac{V\left[E(Y|X_i)\right]}{V(Y)}$$

Total effect index accounts for the total contribution to the output variation due to factor X_1 (if 3 factors), i.e its first-order effect plus all higher-order effects due to interactions

$$S_{T1} = S_1 + S_{12} + S_{13} + S_{123}$$

S_{Ti} give information on the nonadditive features of the model

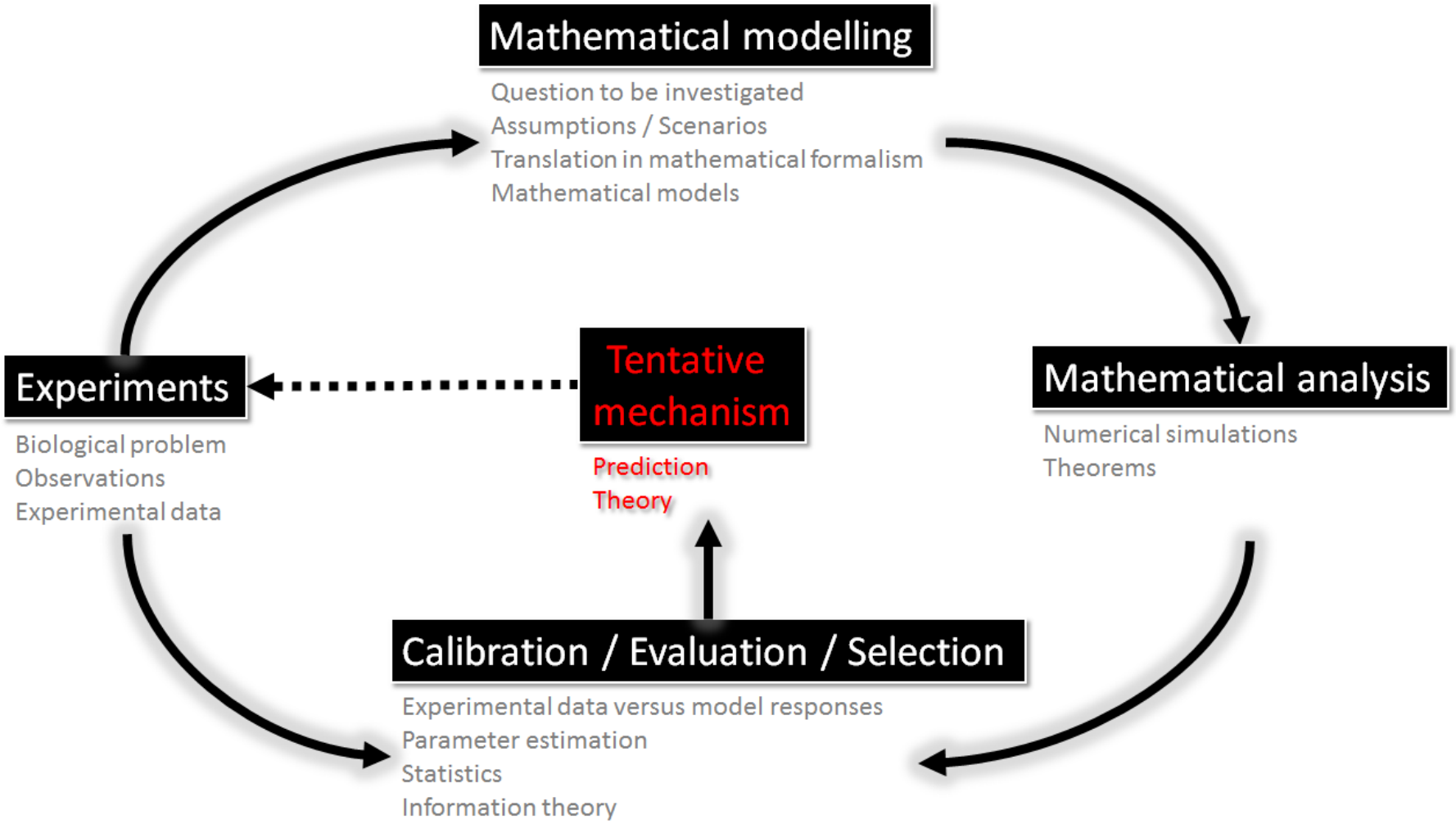
$$S_{Ti} = 1 - \frac{V\left[E(Y|X_{\sim i})\right]}{V(Y)}$$

A significant difference between S_{Ti} and S_i signals important interaction involving that factor

The condition $S_{Ti} = 0$ is necessary and sufficient for X_i to be a non-influential factor

Outline

Work flow – Mathematical modelling



Portet (2015) Insights E-Journal. 8(2).

My $0.02 on combining mathematical modelling and experimental data

To get more: Collection of models (with model selection):

- test different scenarios
- select model(s) that approximate the best data considered (Still not the truth!!)
- evaluate the relative importance of processes

To keep in mind: Conclusions drawn when combining mathematical modelling and data are impacted by

- mathematical translation of biological processes
- data considered

$\Rightarrow$ change in sensitivity analysis outcomes, transient (reactivity) and long term (prediction) dynamics...

Al-Darabsah, K.-L. Liao, and S. Portet, A simple in-host model for COVID-19 with treatments: model prediction and calibration.

Journal of Mathematical Biology, **86**:20 (2023)

Why use mathematical modelling?

- Test a large number of different scenarios
- Propose tentative hypotheses to be tested $\Rightarrow$ propose new experiments
- Identify the major components of processes
- Extrapolate the broad behavior of a system for which data cannot easily be obtained
- Theorize the processes (clarify hypotheses and characterize the chain of events)

“All models are wrong, but some are useful.” G. E. P. Box