Bioprocess modeling and estimation

An introduction to parameter identification

Alain Vande Wouwer[1]   Philippe Bogaerts[2]

[1] Control Department
University of Mons

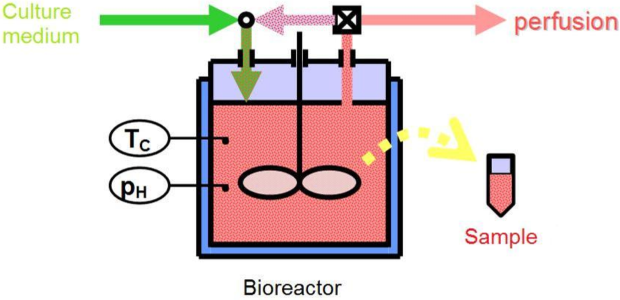
[2] 3BIO
University of Brussels

Graduate School, March 2010

Purpose of the talk : identification of macroscopic model of bioprocesses

- introduce important parameter estimation methods
  - stress the limits of the least-squares approach
  - present the maximum likelihood approach and some useful analysis tools
- present procedures for estimating the yield coefficients and the kinetic parameters in biological models

Culture of microorganisms in bioreactors



Model structure

A reaction scheme:

$$\sum_{i \in R_m} (-k_{i,m})\xi_i \overset{\varphi_m}{\rightarrow} \sum_{j \in P_m} k_{j,m}\xi_j \qquad m \in [1, ..., M]$$

Examples:

- a simple bacterial growth $$k_S S \rightarrow X$$
- culture of S. cerevisiae
$$G + k_5 O \overset{\varphi_1}{\rightarrow} k_1 X + k_7 C$$
$$G \overset{\varphi_2}{\rightarrow} k_2 X + k_4 E + k_8 C$$
$$E + k_6 O \overset{\varphi_3}{\rightarrow} k_3 X + k_9 C$$

The reactions are assumed independent (M is minimum) and the number of components N ≥ M

Model structure

A system of mass balance equations:

$$\frac{d\left(V(t)\xi(t)\right)}{dt} = V(t)K\varphi(\xi,t) - F_{out}(t)\xi(t) + F_{in}(t)\xi_{in}(t)$$

$$\frac{dV(t)}{dt} = F_{in}(t) - F_{out}(t)$$

$$\begin{aligned} V(t)\frac{d\xi(t)}{dt} + \xi(t)\frac{dV(t)}{dt} &= V(t)\frac{d\xi(t)}{dt} + \xi(t)\left(F_{in}(t) - F_{out}(t)\right) \\ &= V(t)K\varphi(\xi,t) - F_{out}(t)\xi(t) + F_{in}(t)\xi_{in}(t) \end{aligned}$$

$$\frac{d\xi(t)}{dt} = K\varphi(\xi,t) + \frac{F_{in}(t)}{V(t)}\left(\xi_{in}(t) - \xi(t)\right) = K\varphi(\xi,t) + D(t)\left(\xi_{in}(t) - \xi(t)\right)$$

Model structure

A general macroscopic model:

$$\frac{d\xi(t)}{dt} = K\varphi(\xi, t) - D(t)\xi(t) + F(t) - Q(t)$$

Examples:

- a simple bacterial growth $$\begin{bmatrix} \dot{X} \\ \dot{S} \end{bmatrix} = \begin{bmatrix} 1 \\ -k_S \end{bmatrix} \varphi(S, X)$$
- culture of S. cerevisiae

$$\begin{bmatrix} \dot{X} \\ \dot{G} \\ \dot{E} \\ \dot{O} \\ \dot{P} \end{bmatrix} = \begin{bmatrix} k_1 & k_2 & k_3 \\ -1 & -1 & 0 \\ 0 & k_4 & -1 \\ -k_5 & 0 & -k_6 \\ k_7 & k_8 & k_9 \end{bmatrix} \cdot \begin{bmatrix} \mu_1 \cdot X \\ \mu_2 \cdot X \\ \mu_3 \cdot X \end{bmatrix} - D \cdot \begin{bmatrix} X \\ G \\ E \\ O \\ P \end{bmatrix} + \begin{bmatrix} 0 \\ G_{in} \cdot D \\ E_{in}.D \\ OTR \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ CTR \end{bmatrix}$$

Model structure

A variety of kinetic laws:

- Monod law $\mu(S) = \mu_{\max} \frac{S}{K_M + S}$
- Contois law $\mu(S) = \mu_{\max} \frac{S}{K_M X + S}$
- Haldane law $\mu(S) = \mu_{\max} \frac{S}{K_M + S + S^2/K_I}$

A wrong model structure will lead to systematic errors in the parameters. Overparametrization (a too large set of the degrees of freedom) will lead to increased standard deviations.

Parameter estimation: an art !

Parameter estimation is a difficult task as it combines:

- Informative experimental studies
- Selection of a good model structure (in the following we will assume that this choice is well made: certainly requires a good physical/biological inspiration !)
- Selection of good parameter estimation tools (the goal of this presentation !)
- Model validation (direct and cross validation)

An example of the importance of model structure and validation

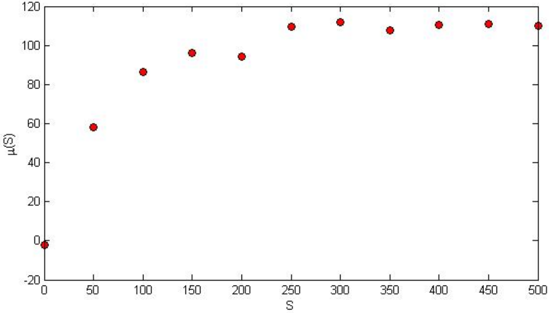
Identification of a kinetic model based on 11 measurements of μ for S from 0 to 500 (by steps of 50) with an additive white noise with normal distribution, zero mean and standard deviation of 5

$$\mu = 120 \frac{S}{40 + S}$$

An example of the importance of model structure and validation

Identification of a kinetic model based on 11 measurements of μ for S from 0 to 500 (by steps of 50) with an additive white noise with normal distribution, zero mean and standard deviation of 5

$$\mu = 120 \frac{S}{40 + S}$$



An example of the importance of model structure and validation

Different candidate model structures

1. Exact model structure (Monod law)

$$\mu = \theta_1 \frac{S}{\theta_2 + S}$$

2. Same level of complexity (Tessier law)

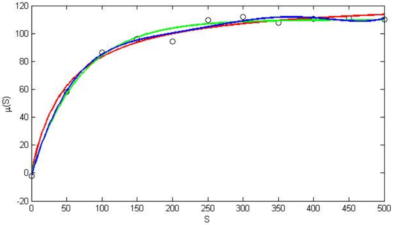
$$\mu = \theta_1 \left(1 - e^{-S/\theta_2}\right)$$

3. Overparametrized model (polynomial ”black-box” model)

$$\mu = \theta_1 + \vartheta_2 S + \theta_3 S^2 + \theta_4 S^3 + \theta_5 S^4 + \theta_6 S^5$$

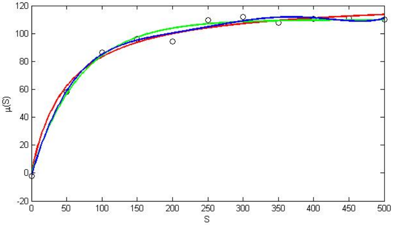
An example of the importance of model structure and validation

Direct Validation

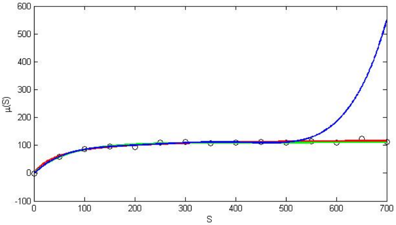


An example of the importance of model structure and validation

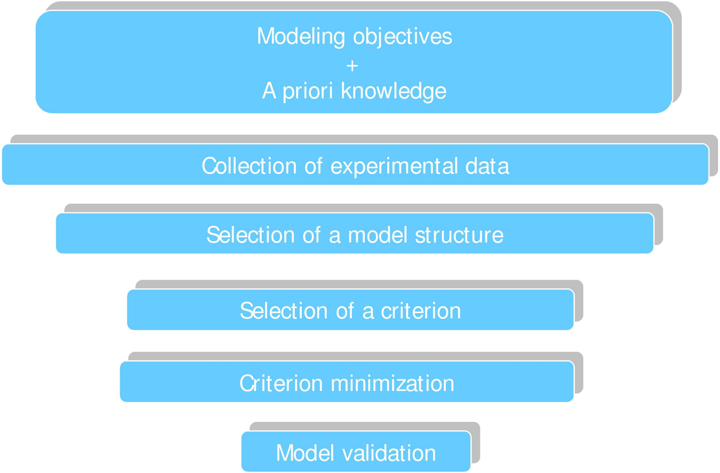
Direct Validation



Cross Validation



Parameter estimation: an art !



Modeling objectives
+
A priori knowledge

Collection of experimental data

Selection of a model structure

Selection of a criterion

Criterion minimization

Model validation

Basic terminology

- Observations (or measurements) are realizations of stochastic processes
- The function of the observations used to compute the parameters is called an estimator.
- Parameters are therefore stochastic variables with an expectation (mean) and a standard deviation
- The deviation of the expectation of the estimator from the hypothetical true value of the parameter is called the bias of the estimator.
- The standard deviation of the estimator defines and quantifies its precision

Advantages of statistical estimation methods

- The possibility to compute accuracy (bias ?) and precision (standard deviation ?)
- The possibility to establish accurate (or unbiased) and precise methods
- The possibility to compute the attainable precision based on a set of available measurements
- The possibility to optimally design experiments (i.e. to design an experiment so as to attain a prescribed precision or to optimize the precision)

Overview of the talk

- Basics of data analysis
- Fisher Information Matrix
- Parameter estimation methods
  - The least-squares approach
  - The maximum likelihood approach
  - Recursive formulations
- Sensitivity analysis
- Identification of the pseudo-stoichiometry and kinetics

Basics of data analysis: distribution of observations

Vector of observations:

$$\underline{y} = [y_1 \cdots y_M]^T$$

for $t = t_1, ..., t_M$, and some unknown parameters $\underline{\theta} = [\theta_1 ... \theta_P]^T$ ....
The purpose of identification is to estimate the **expectation model** of the observations

$$E[y_j] = m(t_j, \underline{\theta})$$

$$\underline{m} = [m_1 \cdots m_M]^T$$

The deviations of the observations y from the values of the expectation model at the measurement points

$$\varepsilon(t_j, \underline{\theta}) = y(t_j) - m(t_j, \underline{\theta})$$

are zero-mean stochastic errors.

Basics of data analysis: distribution of observations

Stochastic variables can be described by deterministic functions:

- Joint probability density function

$$p(\underline{y};\underline{\theta})$$

- Joint log-probability density function

$$q(\underline{y};\underline{\theta}) = \ln(p(\underline{y};\underline{\theta}))$$

- Mean defined through the mathematical expectation

$$m_j = E(y_j) = \int y_j p(y_1, \cdots, y_M; \theta) dy_1 \cdots dy_M$$

- Covariance matrix

$$\Sigma = E[(\underline{y} - \underline{m})(\underline{y} - \underline{m})^T]$$

Basics of data analysis: distribution of observations

The central limit theorem of calculus of probability and the problem of moments [Pólya, 1920]: ”The occurrence of the Gaussian probability density in repeated experiments, in errors of measurements, which result in the combination of very many and very small elementary errors, in diffusion processes etc., can be explained, as is well-known, by the very same limit theorem, which plays a central role in the calculus of probability. The actual discoverer of this limit theorem is to be named Laplace; it is likely that its rigorous proof was first given by Tschebyscheff and its sharpest formulation can be found, as far as I am aware of, in an article by Liapounoff”

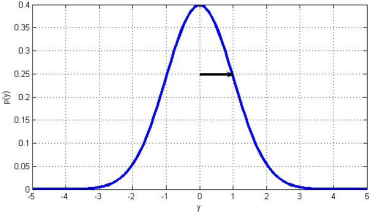
Basics of data analysis: distribution of observations

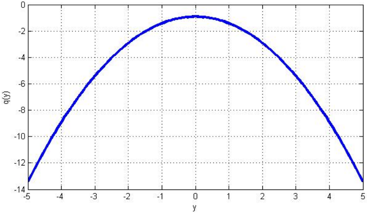
The joint normal probability density function:

$$p(\underline{y}) = \frac{1}{(2\pi)^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(\underline{y}-\underline{m})^T \Sigma^{-1} (\underline{y}-\underline{m})\right)} \equiv N(m, \Sigma)$$

and log-density function:

$$q(\underline{y}) = -\frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (\underline{y} - \underline{m})^T \Sigma^{-1} (\underline{y} - \underline{m})$$





Basics of data analysis: distribution of observations

If the y_j are uncorrelated, $\Sigma = diag(\sigma_1^2 \cdots \sigma_M^2)$ and if all the m_j and σ_j are identical, the y_j are *iid* (independent and identically distributed)

$$q(\underline{y}) = -\frac{M}{2}\ln(2\pi) - \frac{1}{2}M\ln\sigma - \frac{1}{2\sigma^2}\sum_{j=1}^{M}(y_j - m_j)^2$$

If the joint probability density function depends on unknown parameters $\underline{\theta}$ (we assume Σ does not depend on $\underline{\theta}$, i.e. it is known or constant) :

$$p(\underline{y};\underline{\theta}) = \frac{1}{(2\pi)^{\frac{M}{2}}\,|\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(\underline{y}-\underline{m}(\underline{\theta}))^T\Sigma^{-1}(\underline{y}-\underline{m}(\underline{\theta}))\right)}$$

$$q(\underline{y};\underline{\theta}) = -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\underline{y} - \underline{m}(\underline{\theta}))^T\Sigma^{-1}(\underline{y} - \underline{m}(\underline{\theta}))$$

Basics of data analysis: precision and accuracy

Parameters : $\underline{\theta} = [\theta_1 \cdots \theta_P]^T$
Parameter estimates : $\underline{\hat{\theta}} = [\hat{\theta}_1 \cdots \hat{\theta}_P]^T$
Bias: $b_{\theta_k} = E(\hat{\theta}_k) - \theta_k$
Variance: $\mathrm{var}(\hat{\theta}_k) = \sigma^2_{\theta_k} = E(\hat{\theta}_k - E(\hat{\theta}_k))^2$
Mean squared error: $mse(\hat{\theta}_k) = E(\hat{\theta}_k - \theta_k)^2 = \sigma^2_{\theta_k} + b^2_{\theta_k}$

- *An estimator is defined as convergent in quadratic mean if its mean squared error vanishes asymptotically.*
- *An estimator is defined as consistent if the probability that it deviates less than a specified amount from its exact value may be made arbitrarily close to one if the number of observations is taken sufficiently large.*
- *Convergence in quadratic mean implies consistency but the converse is not true.*

Basics of data analysis: precision and accuracy

Other important indicators are the Fisher scores, which are **stochastic variables** as they are evaluated using the actual observations

$$\underline{s}_{\theta} = \frac{\partial q(\underline{y};\underline{\theta})}{\partial \underline{\theta}}$$

Under suitable regularity conditions,

$$E(\underline{s}_{\theta}) = 0$$

For Gaussian distribution (and under the assumption that Σ does not depend on θ)

$$\underline{s}_{\theta} = \frac{\partial m^{T}(\underline{\theta})}{\partial \underline{\theta}} \Sigma^{-1}(\underline{y} - \underline{m}(\underline{\theta})) = \frac{\partial \underline{m}^{T}(\underline{\theta})}{\partial \underline{\theta}} \Sigma^{-1} \underline{\varepsilon}(\underline{\theta})$$

Basics of data analysis: precision and accuracy

$$\underline{s}_\theta = \frac{\partial \underline{m}^T(\theta)}{\partial \underline{\theta}} \Sigma^{-1} \underline{\varepsilon}(\underline{\theta})$$

If the observations are independent :

$$s_{\theta_k} = \sum_{j=1}^{M} \frac{1}{\sigma_j^2} \frac{\partial m_j(\underline{\theta})}{\partial \theta_k} \varepsilon_j(\underline{\theta})$$

$$\mathrm{var}[s_{\theta_k}] = \sum_{j=1}^{M} \frac{1}{\sigma_j^2} \left(\frac{\partial m_j(\underline{\theta})}{\partial \theta_k} \right)^2$$

With each additional observation the variance of the Fisher score increases.

Basics of data analysis: precision and accuracy

The Fisher information matrix :

$$\mathrm{F}_\theta = E[\underline{s}_\theta \underline{s}_\theta^T] = E[\frac{\partial q(\underline{y};\underline{\theta})}{\partial \underline{\theta}} \frac{\partial q(\underline{y};\underline{\theta})}{\partial \underline{\theta}^T}]$$

> *The Fisher information matrix is a covariance matrix, and is therefore positive semidefinite. It is positive definite if and only if the elements of the Fisher score vector are linearly independent stochastic variables.*

The Fisher information matrix of normally distributed observations :

$$\underline{s}_\theta = \frac{\partial \underline{m}^T(\underline{\theta})}{\partial \underline{\theta}} \Sigma^{-1} \underline{\varepsilon}(\underline{\theta})$$

$$\mathrm{F}_\theta = E[\underline{s}_\theta \underline{s}_\theta^T] = \frac{\partial \underline{m}^T(\underline{\theta})}{\partial \underline{\theta}} \Sigma^{-1} \frac{\partial \underline{m}(\underline{\theta})}{\partial \underline{\theta}^T}$$

Basics of data analysis: precision and accuracy

Inflow of Fisher information for independent observations :

$$p(\underline{y};\underline{\theta}) = p_1(y_1;\underline{\theta})p_2(y_2;\underline{\theta})\cdots p_M(y_M;\underline{\theta}) \quad \Rightarrow \quad q(\underline{y};\underline{\theta}) = \sum_{j=1}^{M} q_j(y_j;\underline{\theta})$$

$$s_{\theta_k} = \sum_{j=1}^{M} \frac{\partial q_j(y_j;\underline{\theta})}{\partial \theta_k}$$

$$\mathrm{cov}[s_{\theta_k} s_{\theta_i}] = \sum_{j=1}^{M} E[\frac{\partial q_j(y;\underline{\theta})}{\partial \theta_k}\frac{\partial q_j(y;\underline{\theta})}{\partial \theta_i}] \Rightarrow F_M = \sum_{j=1}^{M} E[\frac{\partial q_j(y;\underline{\theta})}{\partial \underline{\theta}}\frac{\partial q_j(y;\underline{\theta})}{\partial \underline{\theta}^T}]$$

$F_{M+1} \succ F_M$ *: as a result of any additional independent observation, the diagonal elements of the Fisher information matrix either increase or remain the same.*

Basics of data analysis: precision and accuracy

Limits to precision : The Cramer-Rao bound

Suppose that $\hat{\theta} = [\hat{\theta}_1 \cdots \hat{\theta}_P]$ *is an unbiased estimator. Then, under suitable regularity conditions,*

$$\mathrm{cov}(\underline{\hat{\theta}}, \underline{\hat{\theta}}) \succ F_\theta^{-1}$$

- *No unbiased estimator can be found that is more precise than a hypothetical unbiased estimator that has variances equal to the diagonal elements of the Cramer-Rao lower bound matrix.*
- *This inequality does not imply that the off-diagonal elements of the covariance matrix are necessarily larger than or equal to the corresponding elements of the Cramer-Rao lower bound matrix.*

Basics of data analysis: precision and accuracy

Efficiency of an estimator

- *Unbiased estimators attaining the Cramer-Rao lower bound are called efficient unbiased.*
- *Although often no efficient unbiased estimator exists, an estimator can be usually found that attains the Cramer-Rao lower bound asymptotically.*
- *The efficiency of an estimator is the ratio of the relevant Cramer-Rao variance to the mean squared error of the estimator.*
- *A necessary and sufficient condition for an estimator* $\underline{\hat{\theta}}$ *to be unbiased and to attain the Cramer-Rao lower bound is*

$$F_\theta^{-1}\underline{s}_\theta = \underline{\hat{\theta}} - \underline{\theta}$$

Basics of data analysis: precision and accuracy

The Cramer-Rao bound and parameter identifiability

> *A vector of parameters is identifiable from a set of observations if the relevant Fisher information matrix is nonsingular. If, on the other hand, the Fisher information matrix is singular, the vector of parameters is nonidentifiable and the Cramer-Rao lower bound matrix does not exist.*

For normally distributed observations :

$$\mathrm{F}_\theta = \frac{\partial m^T(\underline{\theta})}{\partial \underline{\theta}} \Sigma^{-1} \frac{\partial m(\underline{\theta})}{\partial \underline{\theta}^T}$$

which implies that the parameters of the model are identifiable if and only if both the covariance matrix of the observations and the Jacobian matrix $\frac{\partial m(\underline{\theta})}{\partial \underline{\theta}^T}$ are nonsingular.

Basics of data analysis: precision and accuracy

Two (trivial) examples of singularity of $\frac{\partial m(\underline{\theta})}{\partial \underline{\theta}^T}$:

- the model can be reparametrized with less parameters, i.e. $m_1(\theta_1, \theta_2, ...\theta_{P1})$ can be replaced by $m_2(\theta_1, \theta_2, ...\theta_{P2})$, then

  $$\frac{\partial m_1(\underline{\theta}^{(1)})}{\partial \underline{\theta}^{(1)T}}$$

  is singular and

  $$\frac{\partial m_2(\underline{\theta}^{(2)})}{\partial \underline{\theta}^{(2)T}}$$

  maybe not ...
- the number of observations M is less than the number of parameters P

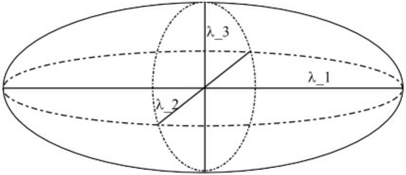
Basics of data analysis: precision and accuracy

The Cramer-Rao bound and experiment design

> *The guiding principle is the assumption that an hypothetical estimator is available that attains the Cramer-Rao lower bound. Experimental design may be used to manipulate the Cramer-Rao variances and even to minimize them in a chosen sense.*

Many criteria have been proposed:

- maximize the determinant of the FIM (volume)
- minimize the condition number of the FIM (ratio of the largest to the smallest axes)
- etc.



Parameter estimation: the linear least-squares

The linear least-squares estimator

$$y(i) = y_m(i) + \varepsilon(i) = \underline{x}^T(i)\underline{\theta} + \varepsilon(i) \quad i = 1, \cdots M$$

$\underline{y} = [y_1 \cdots y_M]$	Measurements (stochastic variables)
$\underline{y}_m$	Model (linear in this case)
$\underline{x} = [x_1 \cdots x_p]$	Explicative variables (perfectly known)
$\underline{\theta} = [\theta_1 \cdots \theta_p]$	Unknown parameters (stochastic variables)
$\underline{\varepsilon}$	Measurement errors (stochastic variables)

Parameter estimation: the linear least-squares

Least-squares criterion

$$J(\underline{\theta}) = \sum_{i=1}^{M} \varepsilon^2(i) = \sum_{i=1}^{M} (y(i) - y_m(i))^2 = \sum_{i=1}^{M} \left(y(i) - \underline{x}^T(i)\underline{\theta}\right)^2$$

Least-squares estimator

$$\underline{\hat{\theta}}_M = \underset{\underline{\theta}}{ArgMin}\ J(\underline{\theta})$$

Necessary condition and solution

$$\left.\frac{\partial J(\underline{\theta})}{\partial \underline{\theta}}\right|_{\underline{\theta}=\underline{\hat{\theta}}_M} = 0$$

$$\underline{\hat{\theta}}_M = \left(\sum_{i=1}^{M} \underline{x}(i)\underline{x}^T(i)\right)^{-1} \sum_{l=1}^{M} \underline{x}(i)y(i) = (X^T X)^{-1} X^T \underline{y}$$

Parameter estimation: the linear least-squares

Properties of the least-squares estimator

Estimation error $\underline{\tilde{\theta}}_M = \underline{\theta} - \underline{\hat{\theta}}_M$

Unbiased estimator

$$E[\underline{\tilde{\theta}}_M] = 0 \quad if \quad E[\varepsilon_i] = 0 \quad \forall i$$

AND if the linear model structure is appropriate (otherwise systematic bias can occur which is not related to the stochastic nature of the signals !)

Covariance

$$E[\underline{\tilde{\theta}}_M \underline{\tilde{\theta}}_M^T] = \sigma^2 \left(\sum_{i=1}^{M} \underline{x}(i)\underline{x}^T(i) \right)^{-1} \quad if \quad E[\varepsilon(i)\varepsilon(j)] = \sigma^2 \delta(i-j) \quad \forall i,j$$

which can be estimated as

$$E[\underline{\tilde{\theta}}_M \underline{\tilde{\theta}}_M^T] = \hat{\sigma}^2 F_\theta^{-1} = \hat{\sigma}^2 \left(\sum^{M} \underline{x}(i)\underline{x}^T(i) \right)^{-1} = \hat{\sigma}^2 (X^T X)^{-1}$$

Parameter estimation: the linear least-squares

How can we estimate $\hat{\sigma}_M^2$? A first natural idea would be

$$\hat{\sigma}_M^2 = \frac{J(\underline{\theta}_M)}{M} = \frac{1}{M}(\underline{y} - X\underline{\hat{\theta}})^T(\underline{y} - X\underline{\hat{\theta}})$$

However this estimator is biased !

$$\underline{y} - X\underline{\theta} = \underline{\hat{\varepsilon}}$$

$$\underline{y} - X\underline{\hat{\theta}}_M = \underline{y} - X\underline{\theta} + X(\underline{\theta} - \underline{\hat{\theta}}_M) = \underline{\hat{\varepsilon}} + X\underline{\tilde{\theta}}_M = (I_M - A)\underline{\hat{\varepsilon}}$$

with $A = X(X^TX)^{-1}X^T$ a symetric matrix with the following properties

$$AA = A \quad AX\underline{\theta} = X\underline{\theta}$$

Parameter estimation: the linear least-squares

$$E[\hat{\sigma}_M^2] = \frac{1}{M} E[\hat{\varepsilon}^T (I_M - A)\hat{\varepsilon}]$$

where

$$E[\underline{\hat{\varepsilon}}^T I_M \underline{\hat{\varepsilon}}] = M\sigma^2 \qquad E[\underline{\hat{\varepsilon}}^T A \underline{\hat{\varepsilon}}] = \sigma^2 trA = P\sigma^2$$

since

$$trA = tr(X(X^T X)^{-1} X^T) = tr(X^T X(X^T X)^{-1}) = tr(I_P) = P$$

Therefore

$$E[\hat{\sigma}_M^2] = (1 - \frac{P}{M})\sigma^2$$

Parameter estimation: the linear least-squares

A correct estimate $\hat{\sigma}_M^2$ is

$$\hat{\sigma}_M^2 = \frac{J(\underline{\theta}_M)}{M - P} = \frac{1}{M - P}(\underline{y} - X\underline{\hat{\theta}})^T(\underline{y} - X\underline{\hat{\theta}})$$

The linear least-squares: an example

Identification of stoichiometry in a batch culture where the following reaction occurs

$$S_1 + 2S_2 \overset{\varphi}{\to} X \qquad k_{S_1} = -1 \quad k_{S_2} = -2$$

The kinetics is known to be

$$\varphi(S_1, S_2) = 0,01 S_1 S_2$$

The mass balance equations are

$$\frac{dS_1}{dt} = k_{S_1}\varphi(S_1, S_2) \quad \frac{dS_2}{dt} = k_{S_2}\varphi(S_1, S_2) \quad \frac{dX}{dt} = \varphi(S_1, S_2)$$

The linear least-squares: an example

The main idea is to eliminate the kinetics

$$\frac{dS_1}{dt} = k_{S_1}\varphi(S_1, S_2) = k_{S_1}\frac{dX}{dt}$$

and to integrate the resulting equation

$$S_1(i) = k_{S_1}X(i) + (S_1(0) - k_{S_1}X(0))$$

so as to get a linear expression, leading to the estimation of k_{S_1}

$$y(i) = \underline{x}^T(i)\underline{\theta}$$

with

$$y(i) = S_1(i) \quad \underline{x}^T(i) = \begin{bmatrix} X(i) & 1 \end{bmatrix} \quad \underline{\theta}^T = \begin{bmatrix} k_{S_1} & (S_1(0) - k_{S_1}X(0)) \end{bmatrix}$$

The linear least-squares: an example

The same can be done so as to estimate k_{S_2}

$$\frac{dS_2}{dt} = k_{S_2}\varphi(S_1, S_2) = k_{S_2}\frac{dX}{dt}$$

$$y(i) = \underline{x}^T(i)\underline{\theta}$$

$$y(i) = S_2(i) \quad \underline{x}^T(i) = \left[\begin{array}{cc} X(i) & 1 \end{array}\right] \quad \underline{\theta}^T = \left[\begin{array}{cc} k_{S_2} & (S_2(0) - k_{S_2}X(0)) \end{array}\right]$$

Two liner least-squares problems can therefore be solved under the assumption that

- X is known accurately (ideally without any error)
- S_1 and S_2 are measured (in this example every hour, with a zero-mean, Gaussian noise, with a standard deviation of 0.2)

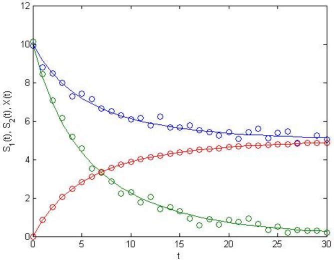
The linear least-squares: an example

Identification of stoichiometry in a batch culture where the following reaction occurs

$$S_1 + 2S_2 \xrightarrow{\varphi} X \qquad k_{S_1} = -1 \quad k_{S_2} = -2$$

MATLAB DEMO

$$\hat{k}_{S_1} = -0.9764 \quad \hat{k}_{S_2} = -2.0466$$

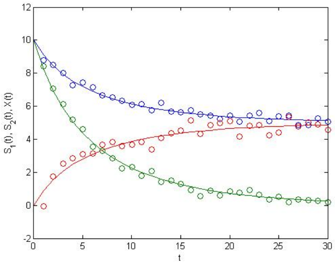


The linear least-squares: an example

If noise is present on X, for instance with a standard deviation of 0.5, a bias appears

MATLAB DEMO

$$\hat{k}_{S_1} = -0.8969 \quad \hat{k}_{S_2} = -1.9145$$



Parameter estimation: the maximum likelihood estimation

The maximum likelihood estimator $\hat{\underline{\theta}}$ from observations y maximizes the the likelihood function $p(y;\underline{\theta})$

$$\hat{\underline{\theta}} = \arg\max_{\theta} p(y;\underline{\theta}) = \arg\max_{\theta} q(y;\underline{\theta})$$

A necessary condition can be expressed using the Fisher score

$$s_{\theta}(\hat{\underline{\theta}}) = \left.\frac{\partial q(y,\underline{\theta})}{\partial \underline{\theta}}\right|_{\underline{\theta}=\hat{\underline{\theta}}} = 0$$

An efficient unbiased estimator is also the maximum likelihood estimator.

Parameter estimation: the maximum likelihood estimation

Additional properties :

- *Under very general conditions, the maximum likelihood estimator is consistent.*
- *Under general conditions, the probability density function of a maximum likelihood estimator tends asymptotically to*

$$\underline{\hat{\theta}} \sim N(\underline{\theta}, F_{\theta}^{-1})$$

Maximum likelihood estimators thus distributed are asymptotically efficient unbiased

Maximum likelihood estimation: the linear case

The linear estimator

$$\begin{aligned} y(i) &= y_m(i) + \varepsilon(i) \\ &= \underline{x}_m^T(i)\underline{\theta} + \varepsilon(i) \\ &= [\underline{x}(i) - \underline{\eta}(i)]^T \underline{\theta} + \varepsilon(i) \quad i = 1, \cdots M \end{aligned}$$

$\underline{y} = [y_1 \cdots y_M]$	Measurements
$\underline{y}_m = [y_{m_1} \cdots y_{m_M}]$	Linear model
$\underline{x}_m = [x_{m_1} \cdots x_{m_P}]$	True explicative variables
$\underline{x} = \underline{x}_m + \underline{\eta}$	Measured explicative variables
$\underline{\theta} = [\theta_1 \cdots \theta_P]$	Unknown parameters
ε and $\underline{\eta}$	Measurement errors

Maximum likelihood estimation: the linear case

The linear estimator

$$y(i) = [\underline{x}(i) - \underline{\eta}(i)]^T \underline{\theta} + \varepsilon(i) \quad i = 1, \cdots M$$

- $E[\underline{\eta}(i)] = 0$
- $E[\underline{\eta}(i)\underline{\eta}^T(j)] = \Sigma(i)\delta(I - j)$
- $p[\underline{\eta}(i)] = \frac{1}{\sqrt{(2\pi)^n |\Sigma(i)|}} \exp\left(-\frac{1}{2}\underline{\eta}^T(i)\Sigma^{-1}(i)\underline{\eta}(i)\right)$
- $E[\varepsilon(i)] = 0$
- $E[\varepsilon(i)\varepsilon(j)] = \sigma^2(i)\delta(i - j)$
- $p[\varepsilon(i)] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\varepsilon^2(i)}{\sigma^2(i)}\right)$
- $E[\varepsilon(i)\underline{\eta}^T(j)] = 0 \quad \forall i, j$

- Zero mean
- White
- Gaussian
- Zero mean
- White
- Gaussian
- uncorrelated

Maximum likelihood estimation: the linear case

Probability of the *M* measurements (assumed independent)

$$p[\underline{\eta}(i), \varepsilon(i); i = 1, ..., M] = \prod_{i=1}^{M} p[\underline{\eta}(i)]p[\varepsilon(i)]$$

$$= \prod_{l=1}^{M} \frac{1}{\sqrt{(2\pi)^n |\Sigma(i)|}} \exp\left(-\frac{1}{2}\underline{\eta}^T(i)\Sigma^{-1}(i)\underline{\eta}(i)\right)\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{\varepsilon^2(i)}{\sigma^2(i)}\right)$$

$$q[\underline{\eta}(i), \varepsilon(i); i = 1, ..., M] = -\frac{1}{2}\sum_{i=1}^{M} \ln\left((2\pi)^n|\Sigma(i)|\right)$$

$$-\frac{1}{2}\sum_{i=1}^{M} \underline{\eta}^T(i)\Sigma^{-1}(i)\underline{\eta}(i) - \frac{1}{2}\sum_{i=1}^{M} \ln\left(2\pi\sigma\right) - \frac{1}{2}\sum_{i=1}^{M} \frac{\varepsilon^2(i)}{\sigma^2(i)}$$

The most likely errors are maximizing this probability, while satisfying the assumed linear model

Maximum likelihood estimation: the linear case

The maximization problem reduces to :

$$\underset{\underline{\eta}(i),\varepsilon(i)}{argmax}\left(\frac{1}{2}\sum_{l=1}^{M}\underline{\eta}^{T}(i)\Sigma^{-1}(i)\underline{\eta}(i)+\frac{1}{2}\sum_{l=1}^{M}\frac{\varepsilon^{2}(l)}{\sigma^{2}(l)}\right)$$

and with the model constraint (Lagrangian formulation)

$$L(\underline{\theta},\underline{y}(i),\underline{x}(i),\lambda(i))=$$
$$\sum_{i=1}^{M}\left(\frac{(y_m(i)-y(i))^2}{\sigma^2(i)}+\left(\underline{x}_m(i)-\underline{x}(l)\right)^T\Sigma^{-1}(i)\left(\underline{x}_m(l)-\underline{x}(i)\right)+2\lambda(i)\left(y(i)-\underline{x}^T(i)\underline{\theta}\right)\right)$$

Maximum likelihood estimation: the linear case

The optimality conditions are

$$\left.\frac{\partial L}{\partial \underline{\theta}}\right|_{\substack{\underline{\theta}=\underline{\hat{\theta}}_M \\ y(i)=\hat{y}_M(i) \\ \underline{x}(i)=\underline{\hat{x}}_M(i) \\ \lambda(i)=\hat{\lambda}_M(i)}} = 0 \qquad \left.\frac{\partial L}{\partial y(i)}\right|_{\substack{\underline{\theta}=\underline{\hat{\theta}}_M \\ y(i)=\hat{y}_k(i) \\ \underline{x}(i)=\underline{\hat{x}}_M(i) \\ \lambda(i)=\hat{\lambda}_M(i)}} = 0$$

$$\left.\frac{\partial L}{\partial \underline{x}(i)}\right|_{\substack{\underline{\theta}=\underline{\hat{\theta}}_M \\ y(i)=\hat{y}_k(i) \\ \underline{x}(i)=\underline{\hat{x}}_M(i) \\ \lambda(i)=\hat{\lambda}_M(i)}} = 0 \qquad \left.\frac{\partial L}{\partial \lambda(i)}\right|_{\substack{\underline{\theta}=\underline{\hat{\theta}}_M \\ y(i)=\hat{y}_k(i) \\ \underline{x}(i)=\underline{\hat{x}}_M(i) \\ \lambda(i)=\hat{\lambda}_M(i)}} = 0$$

Maximum likelihood estimation: the linear case

Using the optimality criterion with respect to $\hat{y}_M(i), \hat{\underline{x}}_M(i), \hat{\lambda}_M(i)$:

$$L(\underline{\theta}) = \frac{1}{2} \sum_{i=1}^{M} \frac{\left(y_m(i) - \underline{x}_m^T(i)\underline{\theta}\right)^2}{\sigma^2(i) + \underline{\theta}^T \Sigma(i) \underline{\theta}}$$

- numerical optimization is required;
- if $\Sigma \equiv 0$, then the maximum likelihood estimator reduces to the least-squares estimator;
- the estimates of $\hat{y}_M(i), \hat{\underline{x}}_M(i), \hat{\lambda}_M(i)$:

$$\begin{aligned}
\hat{y}_M(i) &= y_m(l) - \sigma^2(i)\hat{\lambda}_M(i) \\
\hat{\underline{x}}_M(i) &= \underline{x}_m(i) + \Sigma(i)\hat{\underline{\theta}}_M\hat{\lambda}_M(i) \\
\hat{\lambda}_M(i) &= \frac{y_m(i) - \underline{x}_m^T(i)\hat{\underline{\theta}}_M}{\sigma^2(i) + \hat{\underline{\theta}}_M^T\Sigma(i)\hat{\underline{\theta}}_M}
\end{aligned}$$

Maximum likelihood estimation: the linear case

Statistical properties
... after some calculation neglecting second-order errors

$$\hat{E}[\underline{\tilde{\theta}}_M] \approx 0$$

$$\hat{E}[\underline{\tilde{\theta}}_M \underline{\tilde{\theta}}_M^T] \approx \left(\sum_{i=1}^{M} \frac{\hat{x}_M(i)\hat{x}_M^T(i)}{\sigma^2(i) + \underline{\hat{\theta}}_M^T \Sigma(i) \underline{\hat{\theta}}_M} \right)^{-1}$$

Maximum likelihood estimation: an example

Back to the identification of stoichiometry in a batch culture where the following reaction occurs

$$S_1 + 2S_2 \xrightarrow{\varphi} X \qquad k_{S_1} = -1 \quad k_{S_2} = -2$$

The linear estimator is still given by

$$y(i) = \underline{x}^T(i)\underline{\theta}$$

with

$$y(i) = S_1(i) \quad \underline{x}^T(i) = \left[\begin{array}{cc} X(i) & 1 \end{array} \right] \quad \underline{\theta}^T = \left[\begin{array}{cc} k_{S_1} & (S_1(0) - k_{S_1}X(0)) \end{array} \right]$$

but the measurement errors on the biomass X are taken into account.

MATLAB DEMO

$$\hat{k}_{S_1} = -1.0218 \quad \hat{k}_{S_2} = -2.1005$$

Maximum likelihood estimation: more general cases

Nonlinear (scalar) model

$$y(t_i) = y_m(t_i, \underline{\theta}) + \varepsilon(i) \quad i = 1, \cdots M$$

- $p(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_M) = \prod_{i=1}^{M} p(\varepsilon_i)$
- $e(t_i, \underline{\theta}) = y(t_i) - y_m(t_i, \underline{\theta})$
- $\varepsilon_i \sim N(0, \sigma_i^2)$

$$p(\varepsilon_i) = (2\pi\sigma_i^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\varepsilon_i}{\sigma_i}\right)^2\right\}$$

- Independent additive measurement errors
- output error (= noise $\varepsilon(i)$ for the exact parameters)
- Normal distribution with known variances

Maximum likelihood estimation: more general cases

The joint probability density function is given by:

$$p(\underline{y};\underline{\theta}) = \prod_{i=1}^{M} (2\pi\sigma_i^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{y(t_i) - y_m(t_i,\underline{\theta})}{\sigma_i}\right)^2\right\}$$

and the log-likelihood function

$$q(\underline{y};\underline{\theta}) = -\frac{1}{2}\sum_{i=1}^{M} \ln(2\pi\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{M}\left(\frac{y(t_i) - y_m(t_i,\underline{\theta})}{\sigma_i}\right)^2$$

The maximum likelihood estimator

$$\hat{\underline{\theta}} = \arg\max_{\theta} q(y;\underline{\theta}) = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{M}\left(\frac{y(t_i) - y_m(t_i,\underline{\theta})}{\sigma_i}\right)^2$$

Maximum likelihood estimation: more general cases

Nonlinear (vector) model

$$\underline{y}(t_i) = \underline{y}_m(t_i, \underline{\theta}) + \underline{\varepsilon}(i) \quad i = 1, \cdots M$$

- $p(\underline{\varepsilon}_1, \underline{\varepsilon}_2, \cdots, \underline{\varepsilon}_{n_y}) = \prod_{i=1}^{M} p(\underline{\varepsilon}_i)$
- $\underline{e}(t_i, \underline{\theta}) = \underline{y}(t_i) - \underline{y}_m(t_i, \underline{\theta})$
- $\underline{\varepsilon}_i \sim N(0, \Sigma)$

$$p(\underline{\varepsilon}_i) = ((2\pi)^{n_y} |\Sigma|)^{-1/2} \exp\left\{-\frac{1}{2}\varepsilon_i^T \Sigma^{-1} \varepsilon_i\right\}$$

- Independent errors
- Output errors
- Normal distribution with known covariance matrix

Maximum likelihood estimation: more general cases

The joint probability density

$$p(\underline{y};\underline{\theta}) = \prod_{i=1}^{M} \left((2\pi)^{n_y}|\Sigma|\right)^{-1/2} \exp\left\{\left(\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta})\right)^T \Sigma^{-1} \left(\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta})\right)\right\}$$

The log-likelihood function

$$q(\underline{y};\underline{\theta}) = -\frac{n_y M}{2}\ln 2\pi - \frac{M}{2}\ln\left(|\Sigma|\right) - \frac{1}{2}\sum_{i=1}^{M}\left(\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta})\right)^T \Sigma^{-1} \left(\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta}\right.$$

The maximum likelihood estimator

$$\hat{\underline{\theta}} = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{M}\left(\underline{y}(t_i) - \underline{y}(t_i,\underline{\theta})\right)^T \Sigma^{-1} \left(\underline{y}(t_i) - \underline{y}(t_i,\underline{\theta})\right)$$

Maximum likelihood estimation: more general cases

- If the observations are jointly normally distributed and correlated, the maximum likelihood estimator of the parameters of the expectation model is the weighted least squares estimator with the inverse of the covariance matrix of the observations as weighting matrix.
- If the observations are jointly normally distributed and uncorrelated, the maximum likelihood estimator of the parameters of the expectation model is the least squares estimator with the reciprocals of the variances of the observations as weights.
- If the observations are jointly normally distributed, are uncorrelated, and have equal variances, the maximum likelihood estimator of the parameters of the expectation model is the ordinary least squares estimator.

Maximum likelihood estimation: more general cases

What if the measurement variances are unknown ?
... back to the nonlinear scalar model

$$y(t_i) = y_m(t_i, \underline{\theta}) + \varepsilon(i) \quad i = 1, \cdots M$$

The idea is to extend the vector of unknown parameters

$$\underline{\theta}_{ext} = \left[\begin{array}{cccc} \underline{\theta}^T & \sigma_1 & \cdots & \sigma_M \end{array} \right]^T$$

However, this configuration is non identifiable since $P + M > M$. A reduced parametrization is needed, as for instance,

$$\sigma_i^2(a, b, \vartheta) = a \left| y_m(t_i, \underline{\theta}) \right|^b$$

$$\underline{\theta}_{ext} = \left[\begin{array}{ccc} \underline{\theta}^T & a & b \end{array} \right]^T$$

Maximum likelihood estimation: more general cases

Joint probability density function

$$p(\underline{y};\underline{\theta},a,b)=\prod_{i=1}^{M}\left(2\pi a\left|y_m(t_i,\underline{\theta})\right|^{b}\right)^{-1/2}\exp\left\{-\frac{1}{2}\frac{(\underline{y}(t_i)-\underline{y}_m(t_i,\underline{\theta}))^2}{a\left|\underline{y}_m(t_i,\underline{\theta})\right|^{b}}\right\}$$

Log-likelihood function

$$q(\underline{y};\underline{\theta},a,b)=-\frac{M}{2}\ln 2\pi-\frac{M}{2}\ln a-\frac{b}{2}\sum_{i=1}^{M}\ln\left|\underline{y}_m(t_i,\underline{\theta})\right|-\frac{1}{2a}\sum_{i=1}^{M}\frac{(\underline{y}(t_i)-\underline{y}_m(t_i,\underline{\theta}))^2}{\left|\underline{y}_m(t_i,\underline{\theta})\right|^{b}}$$

which simplifies to the function to minimize

$$J(\underline{\theta},a,b)=M\ln a+b\sum_{i=1}^{M}\ln\left|\underline{y}_m(t_i,\underline{\theta})\right|+\frac{1}{a}\sum_{i=1}^{M}\frac{(\underline{y}(t_i)-\underline{y}_m(t_i,\underline{\theta}))^2}{\left|\underline{y}_m(t_i,\underline{\theta})\right|^{b}}$$

Maximum likelihood estimation: more general cases

Optimality condition

$$\left.\frac{\partial J(\underline{\theta}, a, b)}{\partial a}\right|_{\hat{\underline{\theta}},\hat{a},\hat{b}} = 0 \rightarrow \hat{a} = \frac{1}{M}\sum_{i=1}^{n_t} \frac{(\underline{y}(t_i) - \underline{y}_m(t_i, \hat{\underline{\theta}}))^2}{\left|\underline{y}_m(t_i, \hat{\underline{\theta}})\right|^b}$$

and the criterion simplifies to

$$J(\underline{\theta}, b) = M \ln\left\{\frac{1}{M}\sum_{i=1}^{M} \frac{(\underline{y}(t_i) - \underline{y}_m(t_i, \hat{\underline{\theta}}))^2}{\left|\underline{y}_m(t_i, \hat{\underline{\theta}})\right|^b}\right\} + b \sum_{i=1}^{M} \ln\left|\underline{y}_m(t_i, \underline{\theta})\right|$$

Maximum likelihood estimation: more general cases

In the case of the nonlinear vector model

$$q(y;\underline{\theta},\Sigma)=-\frac{M}{2}\ln((2\pi)^{n_y}|\Sigma|)-\frac{1}{2}\sum_{i=1}^{M}\left(\underline{y}(t_i)-\underline{y}_m(t_i,\underline{\theta})\right)^T\Sigma^{-1}\left(\underline{y}(t_i)-\underline{y}_m(t_i,\underline{\theta})\right)$$

Using the optimality condition

$$\left.\frac{\partial q(y;\underline{\theta},\Sigma)}{\partial\Sigma}\right|_{\hat{\underline{\theta}},\hat{\Sigma}}=0$$

and the relations

$$\frac{\partial\ln|\Sigma|}{\partial\Sigma}=\Sigma^{-1}\qquad\qquad\frac{\partial A\Sigma^{-1}B}{\partial\Sigma}=-\Sigma^{-1}BA\Sigma^{-1}$$

Maximum likelihood estimation: more general cases

The function to minimize is

$$J(\underline{\theta}) = \ln \left| \sum_{i=1}^{M} \left(\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\theta})\right)\left(\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\theta})\right)^T \right|$$

and a a posteriori estimate of the covariance matrix is given by

$$\hat{\Sigma} = \frac{1}{M} \sum_{i=1}^{M} \left(\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\hat{\theta}})\right)\left(\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\hat{\theta}})\right)^T$$

However this estimate, which is biased (as seen before), can be replaced by

$$\hat{\Sigma} = \frac{1}{M-P} \sum_{i=1}^{M} \left(\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\hat{\theta}})\right)\left(\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\hat{\theta}})\right)^T$$

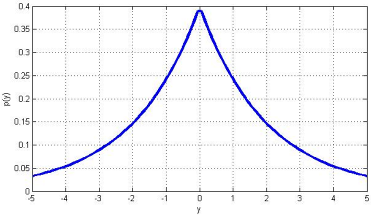
Maximum likelihood estimation: more general cases

What if the measurement noise is not Gaussian ?
... back to the nonlinear scalar model with a Laplacian noise

$$y(t_i) = y_m(t_i, \underline{\theta}) + \varepsilon(i) \quad i = 1, \cdots M$$

with

$$p(\varepsilon_i) = \frac{1}{\sqrt{2}\sigma_i} \exp\left\{ -\frac{\sqrt{2}\,|\varepsilon_i|}{\sigma_i} \right\}$$



Maximum likelihood estimation: more general cases

$$q(y,\underline{\theta}) = -\frac{1}{2}\sum_{i=1}^{M}\ln(2\sigma_i^2) - \sqrt{2}\sum_{i=1}^{M}\frac{\left|\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta})\right|}{\sigma_i}$$

and the criterion to minimize is given by

$$J(\underline{\theta}) = \sum_{i=1}^{M}\frac{\left|\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta})\right|}{\sigma_i}$$

which is the Least Absolute Deviations or Least Absolute Value criterion ... this criterion is useful when there are outliers in the data (robust estimation)

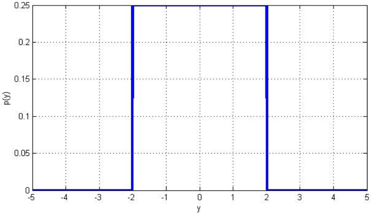
Maximum likelihood estimation: more general cases

... or back to the nonlinear scalar model with a uniform noise

$$y(t_i) = y_m(t_i, \underline{\theta}) + \varepsilon(i) \quad i = 1, \cdots M$$

with

$$p(\varepsilon_i) = \begin{cases} 1/(2a) & if\, |\varepsilon_i| \leq a \\ 0 & otherwise \end{cases}$$



Maximum likelihood estimation: more general cases

The joint probability function

$$p(y, \underline{\theta}) = \begin{cases} 1/(2a)^M \quad \textit{if} \quad \left|\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\theta})\right| \leq a, \quad i = 1, \cdots, M \\ 0 \quad \textit{otherwise} \end{cases}$$

All solutions such that

$$\left|\underline{y}(t_i) - \underline{y}_m(t_i, \hat{\underline{\theta}})\right| \leq a, \quad i = 1, \cdots, M$$

are maximum likelihood estimators

A particular solution is given by the minimax (smallest maximum risk) estimator

$$\hat{\underline{\theta}} = \arg \min_{\underline{\theta}} \max_{1 \leq i \leq M} \left|\underline{y}(t_i) - \underline{y}_m(t_i, \underline{\theta})\right|$$

Maximum likelihood estimation: more general cases

The joint probability function

$$p(y,\underline{\theta}) = \begin{cases} 1/(2a)^M \quad \textit{if} \quad \left|\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta})\right| \leq a, \quad i = 1,\cdots,M \\ 0 \quad \textit{otherwise} \end{cases}$$

All solutions such that

$$\left|\underline{y}(t_i) - \underline{y}_m(t_i,\hat{\underline{\theta}})\right| \leq a, \quad i = 1,\cdots,M$$

are maximum likelihood estimators

A particular solution is given by the minimax (smallest maximum risk) estimator

$$\hat{\underline{\theta}} = \arg\min_{\underline{\theta}} \max_{1\leq i\leq M} \left|\underline{y}(t_i) - \underline{y}_m(t_i,\underline{\theta})\right|$$

Parameter estimation: the nonlinear least-squares

The nonlinear least-squares (scalar) estimator

$$y(i) = y_m(i) + \varepsilon(i) = y(\underline{x}(i), \underline{\theta}) + \varepsilon(i) \quad i = 1, \cdots M$$

$\underline{y} = [y_1 \cdots y_M]$	Measurements (stochastic variables)
$\underline{y}_m$	Model (general nonlinear expression)
$\underline{x} = [x_1 \cdots x_{n_x}]$	Explicative variables (perfectly known)
$\underline{\theta} = [\theta_1 \cdots \theta_p]$	Unknown parameters (stochastic variables)
$\underline{\varepsilon}$	Measurement errors (stochastic variables)

Parameter estimation: the nonlinear least-squares

The criterion

$$J(\underline{\theta}) = \sum_{i=1}^{M} \varepsilon^2(i) = \sum_{i=1}^{M} \left(\underline{y}(t_i) - \underline{y}_m(\underline{x}(i), \underline{\theta})\right)^2$$

and the parameter estimates

$$\hat{\underline{\theta}} = \arg \underset{\underline{\theta}}{min} \, J(\underline{\theta})$$

This problem cannot be solved in closed form and a numerical optimization procedure is required

- Gauss-Newton method
- Levenberg-Marquardt method (e.g. lsqnonlin in MATLAB)

Parameter estimation: the nonlinear least-squares

Parametric sensitivities

$$y_\theta(\underline{x}(i), \hat{\underline{\theta}}) = \left. \frac{\partial y_m(\underline{x}(i), \underline{\theta})}{\partial \underline{\theta}} \right|_{\underline{\theta}=\hat{\underline{\theta}}}$$

Fisher information matrix

$$F_M = \sum_{i=1}^{M} y_\theta(\underline{x}(i), \underline{\theta}) y_\theta^T(\underline{x}(i), \underline{\theta})$$

Statistical properties (first-order approximation)

$$E[\tilde{\underline{\theta}}] \approx F_M^{-1} \sum_{i=1}^{M} E[\varepsilon(i)] y_\theta(\underline{x}(i), \underline{\theta})$$

- $E[\tilde{\underline{\theta}}] \approx 0 \quad$ if $\quad E[\varepsilon(i)] = 0 \quad \forall i$
- $\lim_{M\to\infty} E[\tilde{\underline{\theta}}] = 0 \quad$ since $\quad \lim_{M\to\infty} F_M^{-1} = 0$

Parameter estimation: the nonlinear least-squares

Covariance of the parameter errors

$$P_M = E[\underline{\tilde{\theta}\tilde{\theta}}^T] \approx \sigma^2 F_M^{-1}$$

if $E[\varepsilon(i)\varepsilon(j)] = \sigma^2\delta(i-j) \quad \forall i, j$

A correct estimate of the measurement variance can be obtained as

$$\hat{\sigma}_M^2 = \frac{J(\underline{\hat{\theta}})}{M-p} = \frac{1}{M-p}\sum_{i=1}^{M}\hat{\varepsilon}_M^2(i)$$

Parameter estimation: the nonlinear least-squares

Extension to the vector case

$$\underline{y}(i) = \underline{y}(\underline{x}(i), \underline{\theta}) + \underline{\varepsilon}(i) \quad i = 1, \cdots M$$

$$J(\underline{\theta}) = \sum_{i=1}^{M} (y(i) - y_m(\underline{x}(i), \underline{\theta}))^T \, Q^{-1}(i) \, (y(i) - y_m(\underline{x}(i), \underline{\theta}))$$

- Weighted least-squares : Q is a positive semi-definite matrix
- Maximum likelihood (Gaussian distribution): Q is the covariance matrix of the measurement noise
- Numerical optimization $\underline{\hat{\theta}} = \underset{\underline{\theta}}{argmin}\, J(\underline{\theta})$

Parameter estimation: the nonlinear least-squares

Extension to differential equation model

$$\frac{d\underline{y}(t)}{dt} = \underline{f}(\underline{y}(t), \underline{x}(t), \underline{\theta})$$

come back to the previous case trough numerical integration

$$\underline{y}(t) = \underline{y}(\underline{x}(t), \underline{\theta}, \underline{y}(0))$$

The model is expressed as

$$\underline{y}(i) = \underline{y}_m(\underline{x}(i), \underline{\theta}) + \underline{\varepsilon}(i)$$

which is an algebraic equation where $\underline{y}(0)$ is included either in $\underline{x}(i)$ (if it is known) or $\underline{\theta}$ (if it is unknown)

Parameter estimation: the nonlinear least-squares

Computation of the parametric sensitivities through the solution of a new differential equation system

$$\frac{\partial}{\partial t}\frac{\partial \underline{y}(\underline{x}(t),\underline{\theta},\underline{y}(0))}{\partial \underline{\theta}} = \frac{\partial \underline{f}(\underline{y}(t),\underline{x}(t),\underline{\theta})}{\partial \underline{y}(t)}\frac{\partial \underline{y}(\underline{x}(t),\underline{\theta},\underline{y}(0))}{\partial \underline{\theta}} + \frac{\partial \underline{f}(\underline{y}(t),\underline{x}(t),\underline{\theta})}{\partial \underline{\theta}}$$

or in short form

$$\frac{\partial \underline{y}_{\theta}}{\partial t} = \frac{\partial \underline{f}}{\partial \underline{y}}\underline{y}_{\theta} + \frac{\partial \underline{f}}{\partial \underline{\theta}}$$

with initial condition

$$\underline{y}_{\theta}(t=0) = \frac{\partial \underline{y}(0)}{\partial \underline{\theta}}$$

which is equal to 0 or 1 whether $\underline{y}(0)$ is known or not

Parameter estimation: the nonlinear least-squares

Covariance of the parameter errors

$$E[\underline{\tilde{\theta}\tilde{\theta}}^T] \approx \hat{P}_M = \hat{\sigma}^2 \hat{F}_M^{-1}$$

$$\hat{F}_M = \sum_{i=1}^{M} y_\theta(\underline{x}(i), \underline{\hat{\theta}}) Q^{-1}(i) y_\theta^T(\underline{x}(i), \underline{\hat{\theta}})$$

$$E[\varepsilon(i)\varepsilon(j)] = \sigma^2 \delta(i-j) \quad \forall i, j$$

$$\hat{\sigma}_M^2 = \frac{J(\underline{\hat{\theta}})}{n_y M - p}$$

Nonlinear least-squares: an example

Identification of the parameters of Monod kinetics in a batch culture

$$S \xrightarrow{\varphi} X$$

$$\varphi(S, X) = \mu(S)X = \mu_{\max} \frac{S}{K + S} X$$

with $\mu_{\max} = 0.2$ and $K = 1$

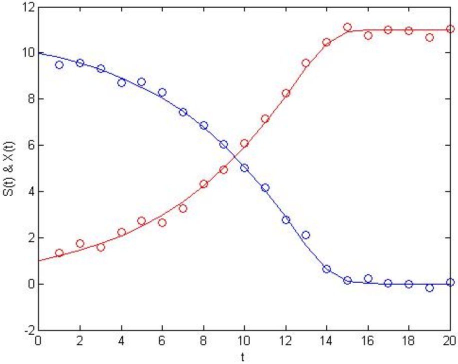
$$\frac{dS}{dt} = -\varphi(S, X) \qquad \frac{dX}{dt} = \varphi(S, X)$$

Measurement of S and X every hour, with a zero-mean Gaussian noise with standard deviation 0.2

Nonlinear least-squares: an example

MATLAB DEMO

$$\hat{\mu}_{\max} = 0.1958 \quad \hat{K} = 0.8929$$



Least-squares: recursive forms and evolutionary problems

On-line estimation of the parameters: recursive formulation of the least-squares algorithm (scalar linear case)

$$
\begin{aligned}
\hat{\underline{\theta}}_{i+1} &= \hat{\underline{\theta}}_i + \frac{P_i \underline{x}(i+1)}{1+\underline{x}^T(i+1) P_k \underline{x}(i+1)} [y(i+1) - \underline{x}^T(i+1)\hat{\underline{\theta}}_i] \\
P_{i+1} &= P_i - \frac{P_i \underline{x}(i+1)\underline{x}^T(i+1) P_i}{1+\underline{x}^T(i+1) P_i \underline{x}(i+1)}
\end{aligned}
$$

- requires an initialization with $\underline{\theta}_0$ and P_0
- usually $P_0 = \lambda I$ is taken large
- P_i is the inverse of the FIM and $P_i \geq P_{i+1}$
- influence of initialization

$$
J'(\underline{\theta}) = \sum_{i=1}^{M} \left(y(i) - \underline{x}^T(i)\underline{\theta} \right)^2 + [\underline{\theta} - \underline{\theta}_0]^T P_0^{-1} [\underline{\theta} - \underline{\theta}_0]
$$

Least-squares: recursive forms and evolutionary problems

Advantages of the recursive formulation

- The number of numerical operations associated with including each new observation is reduced.
- The recursive computation requires a very small and constant amount of memory.
- The recursive estimation and the collection of observations may be stopped once a desired degree of convergence of the parameter estimates has been attained.
- Time-varying parameters can be estimated.

Least-squares: recursive forms and evolutionary problems

An exponential forgetting factor $\alpha \leq 1$ can be included,

$$\begin{aligned}
\hat{\underline{\theta}}_{i+1} &= \hat{\underline{\theta}}_i + \frac{P_i \underline{x}(i+1)}{\alpha + \underline{x}^T(i+1) P_k \underline{x}(i+1)} [y(i+1) - \underline{x}^T(i+1)\hat{\underline{\theta}}_i] \\
P_{i+1} &= \frac{1}{\alpha}[P_i - \frac{P_i \underline{x}(i+1) \underline{x}^T(i+1) P_i}{\alpha + \underline{x}^T(i+1) P_i \underline{x}(i+1)}]
\end{aligned}$$

Scalar nonlinear case:

$$\begin{aligned}
\hat{\underline{\theta}}_{i+1} &= \hat{\underline{\theta}}_k + \frac{P_i \underline{y}_\theta(i+1)}{1 + \underline{y}_\theta^T(i+1) P_i \underline{y}_\theta(i+1)} [y(i+1) - \underline{y}_m(\underline{x}(i+1), \hat{\underline{\theta}}_i)] \\
P_{i+1} &= P_i - \frac{P_i \underline{y}_\theta(i+1) \underline{y}_\theta^T(i+1) P_i}{1 + \underline{y}_\theta^T(i+1) P_i \underline{y}_\theta(i+1)}
\end{aligned}$$

No guarantee of convergence !

recursive forms and evolutionary problems : an example

Identification of stoichiometry in a batch culture where the following reaction occurs

$$S_1 + 2S_2 \xrightarrow{\varphi} X \qquad k_{S_1} = -1 \quad k_{S_2} = -2$$

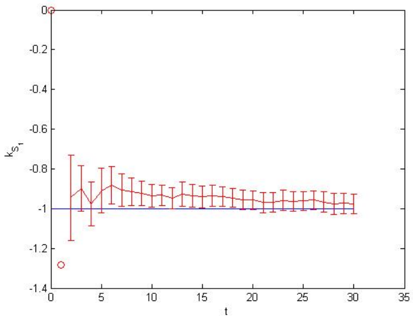
Again, two liner least-squares problems are solved under the assumption that

- X is known accurately (ideally without any error)
- S_1 and S_2 are measured (in this example every hour, with a zero-mean, Gaussian noise, with a standard deviation of 0.2)

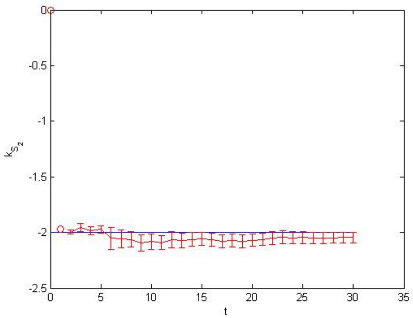
recursive forms and evolutionary problems : an example

Identification of stoichiometry in a batch culture where the following reaction occurs

$$S_1 + 2S_2 \xrightarrow{\varphi} X \qquad k_{S_1} = -1 \quad k_{S_2} = -2$$



0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
k_{S_1}
0
5
10
15
20
25
30
35
t



0
-0.5
-1
-1.5
-2
-2.5
k_{S_2}
0
5
10
15
20
25
30
35
t