Estimation, Model Selection and Optimal Design in Mixed Effects Models Applications to pharmacometrics

Marc Lavielle[1]

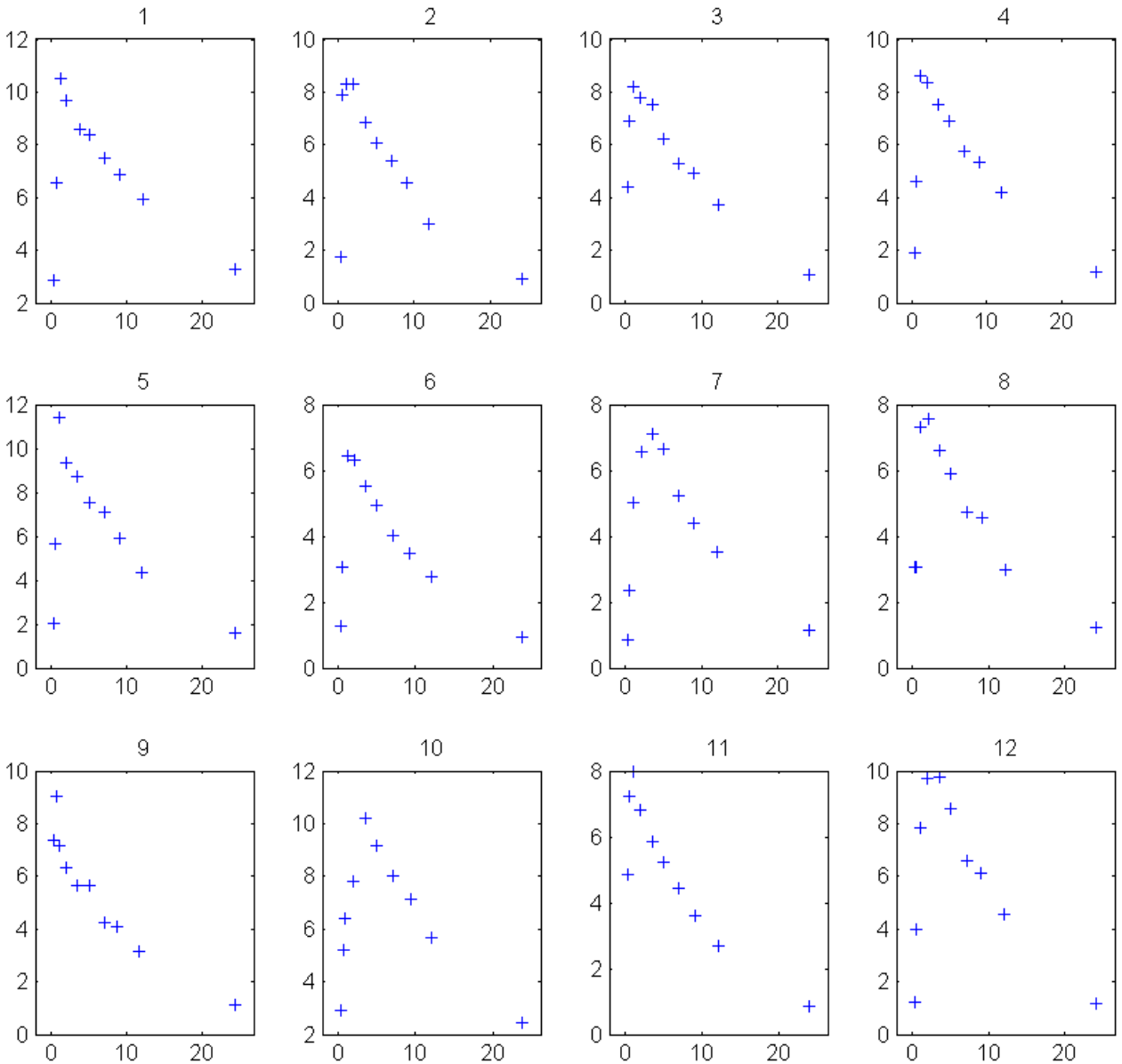
[1]INRIA Saclay

Cemracs 2009 - CIRM

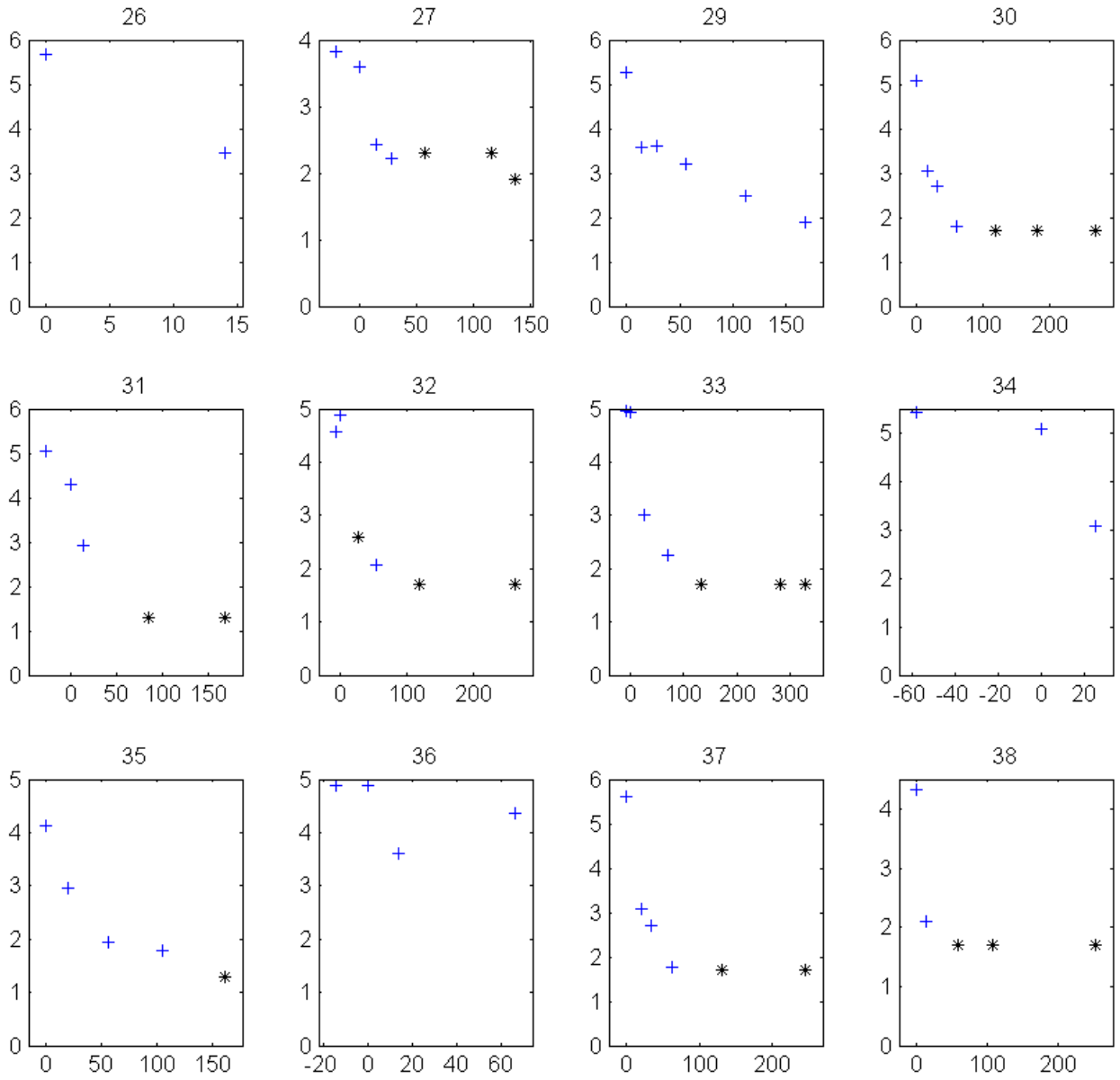
Outline

Some examples of data

Pharmacokinetics of theophylline

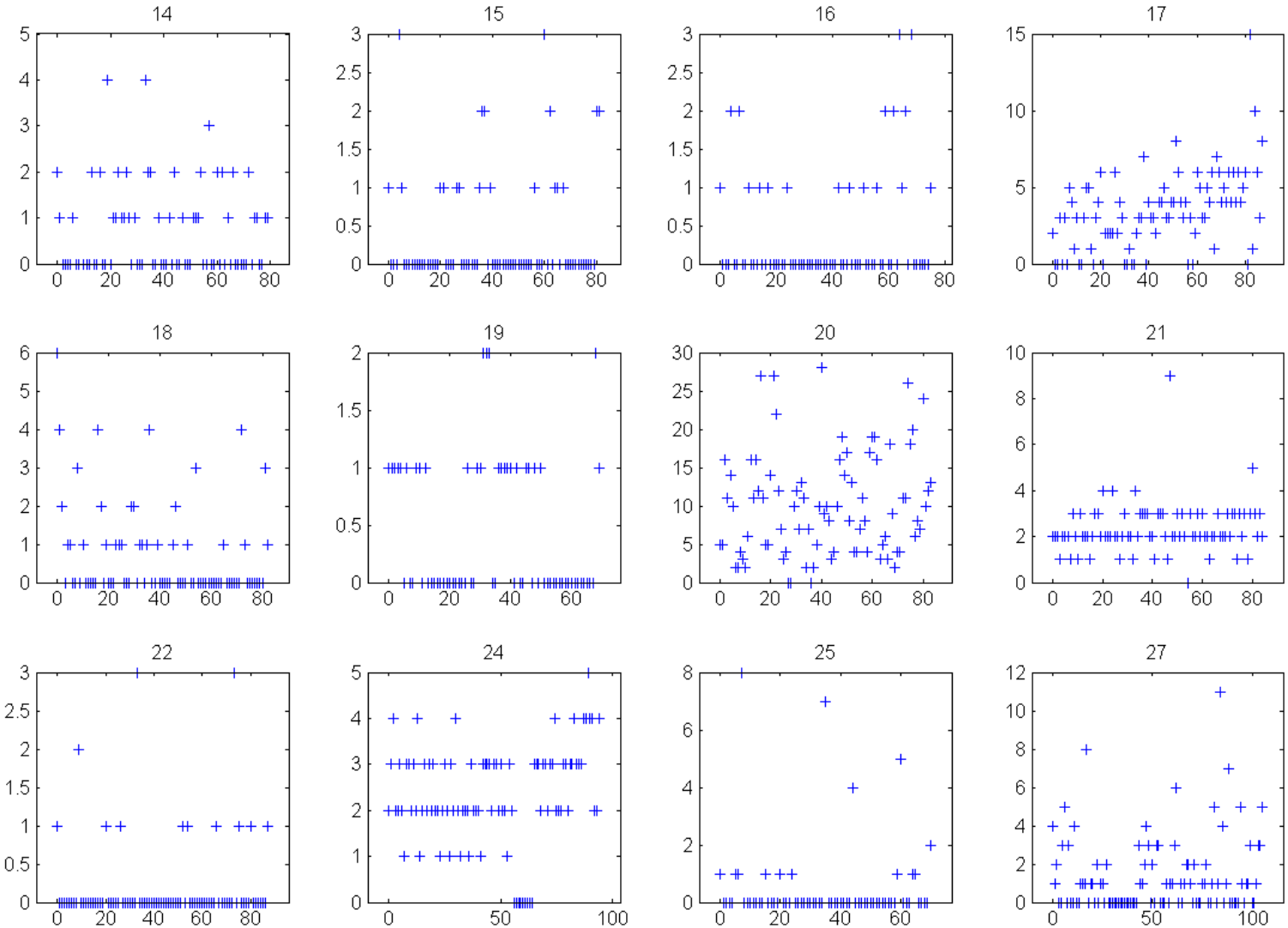


Some examples of data
Viral loads (HIV)



Some examples of data

Daily seizure counts (epilepsy)



The mixed effects model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

$y_{ij} \in \mathbb{R}$ is the jth observation of subject i,

N is the number of subjects

n_i is the number of observations of subject i.

The regression variables, or design variables, (x_{ij}) are **known**,

The individual parameters (ψ_i) are **unknown**.

The mixed effects model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

$y_{ij} \in \mathbb{R}$ is the jth observation of subject i,

N is the number of subjects

n_i is the number of observations of subject i.

The regression variables, or design variables, (x_{ij}) are **known**,

The individual parameters (ψ_i) are **unknown**.

The mixed effects model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

$y_{ij} \in \mathbb{R}$ is the jth observation of subject i,

N is the number of subjects

n_i is the number of observations of subject i.

The regression variables, or design variables, (x_{ij}) are **known**,

The individual parameters (ψ_i) are **unknown**.

The mixed effects model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

The (ψ_i) and the (ε_{ij}) are modelized as sequences of *random variables*.

The goal of the modeler is to develop simultaneously two kinds of models:

(1) The structural model f

(2) The statistical model

The mixed effects model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

(1) **The structural model f:** We are not interested with a purely *descriptive* model which nicely fits the data, but rather with a *mechanistic* model which has some biological meaning and which is a function of some physiological parameters.

Examples:

- compartimental PK models,
- viral dynamic models,
- ...

The mixed effects model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

(2) **The statistical model** aims to explain the variability observed in the data:

- the residual error model: distribution of (ε_{ij})
- the model of the individual parameters: distribution of (ψ_i)

$$\psi_i = h(C_i, \beta, \eta_i)$$

C_i is a vector of covariates
β is a vector of fixed effects
η_i is a vector of random effects

The mixed effects model

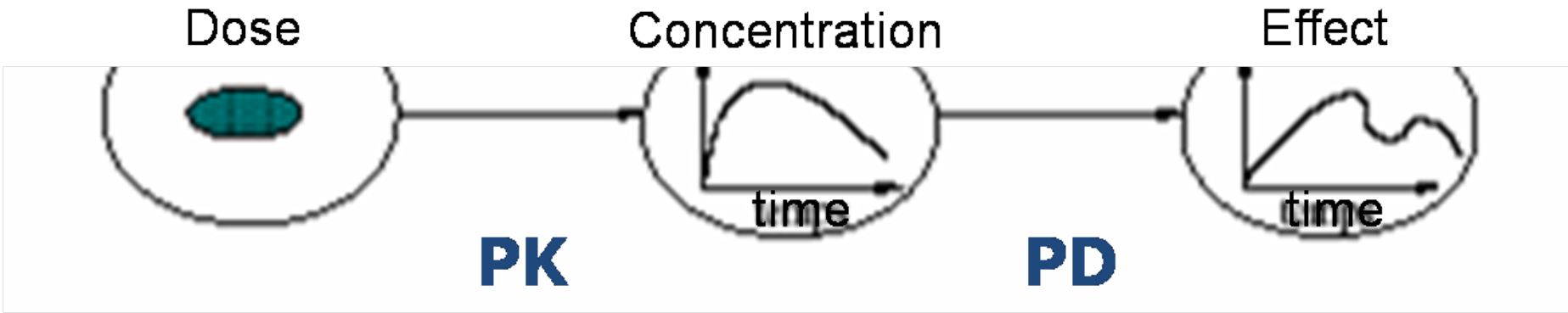
$$
\begin{aligned}
y_{ij} &= f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \\
\psi_i &= h(C_i, \beta, \eta_i)
\end{aligned}
$$

Some statistical issues:

- **Estimation:**
  - estimate the population parameters of the model
  - estimate the individual parameters
- **Model selection and model assessment:**
  - Select and assess the "best" structural model f,
  - Select and assess the "best" statistical model
- **Optimization of the design :**
  - Find the *optimal design* (x_{ij})

Outline

Pharmacokinetics and Pharmacodynamics (PK/PD)

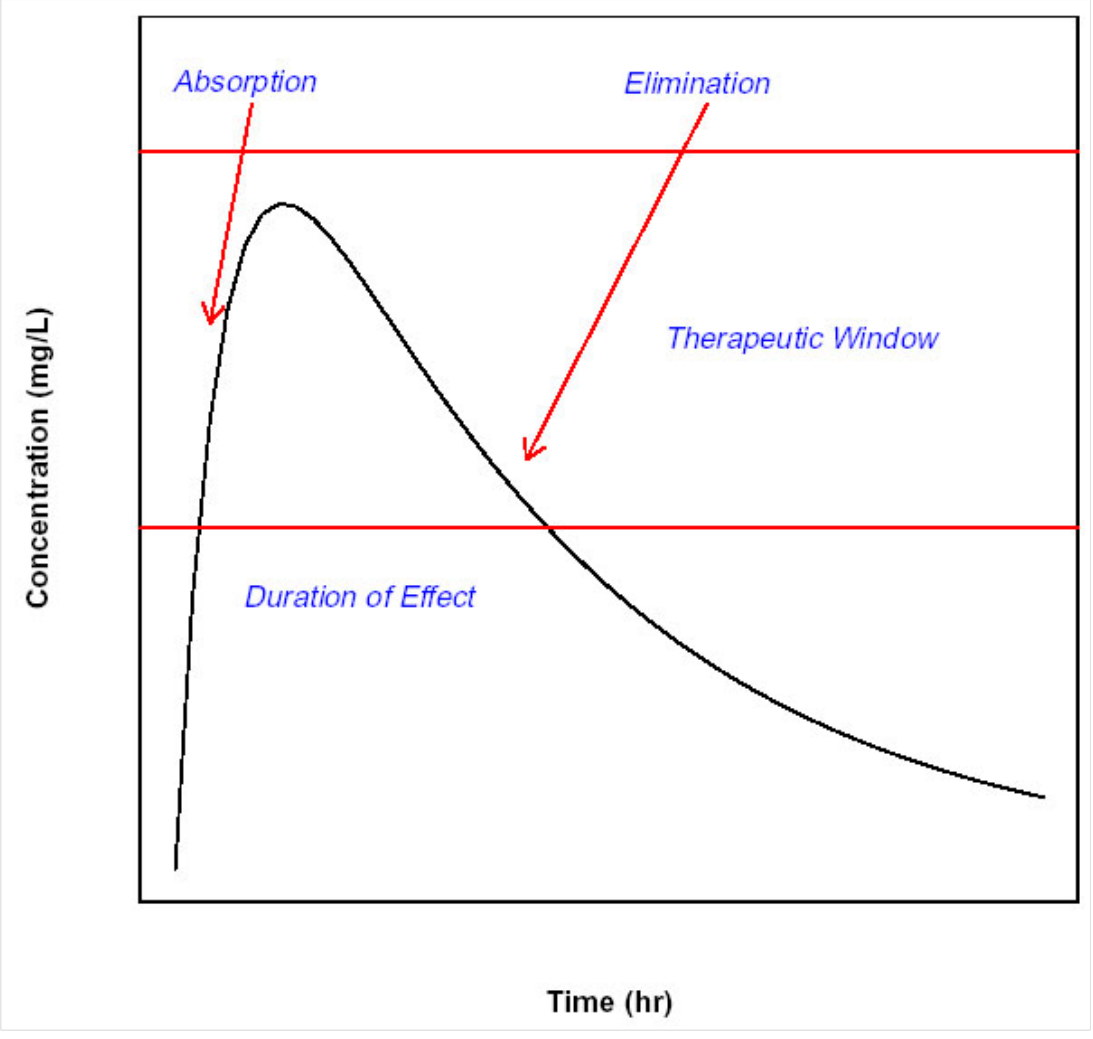


- Pharmacokinetics (PK): "What the body does to the drug"
- Pharmacodynamics (PD): "What the drug does to the body"

Pharmacokinetics and Pharmacodynamics (PK/PD)

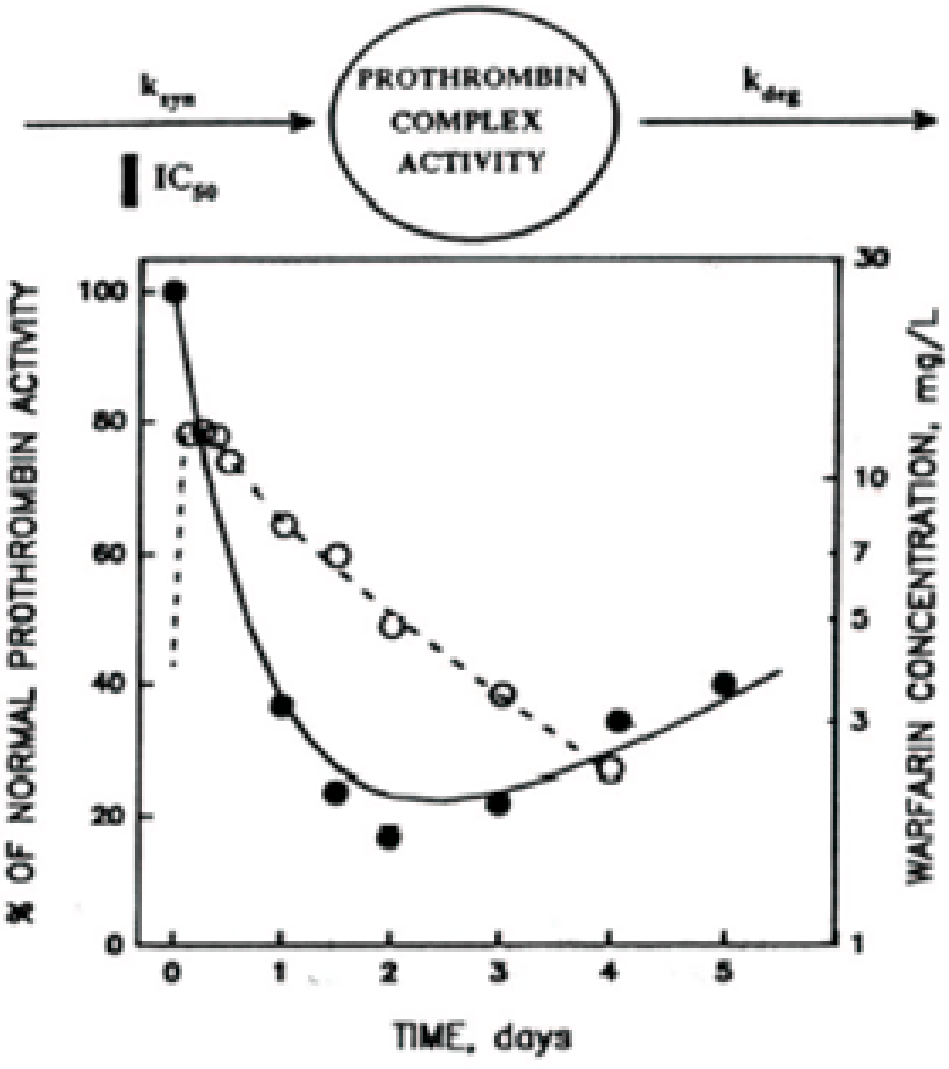
The therapeutic window

Concentrations must be kept high enough to produce a desirable response, but low enough to avoid toxicity.



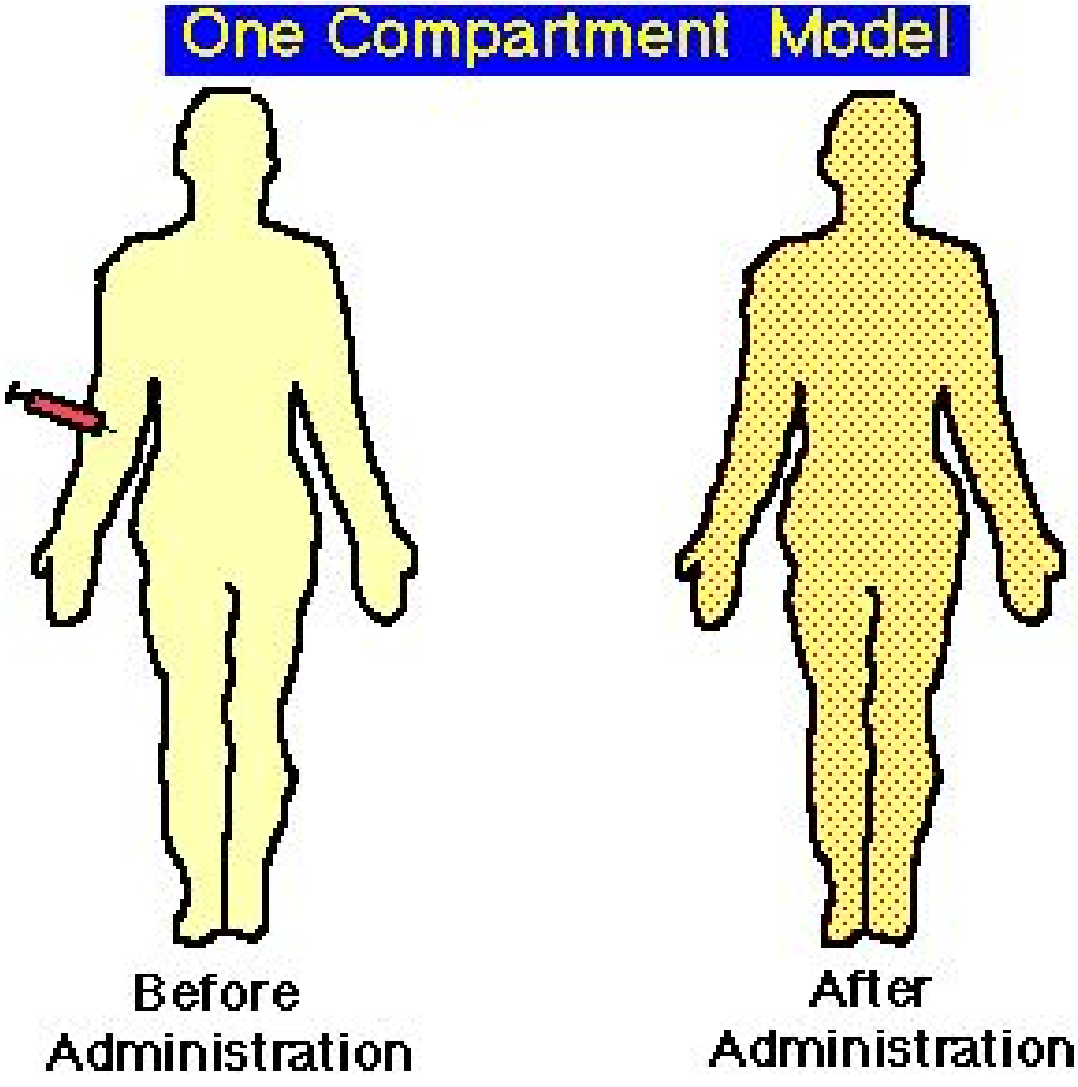
Pharmacokinetics and Pharmacodynamics (PK/PD)

An example PKPD data



Jusko et Ko, *Clin Pharmacol Ther 94*

One compartment PK model
intravenous administration



intravenous administration and first-order elimination

dose D (t=0) $\longrightarrow$ DRUG AMOUNT $Q(t)$ $\longrightarrow$ elimination (rate k_e)

$$\frac{dQ}{dt}(t) = -kQ(t) \quad ; \quad Q(0) = D$$

$$\begin{aligned} Q(t) &= De^{-kt} \\ C(t) &= \frac{Q(t)}{V} = \frac{D}{V}e^{-k_e t} \end{aligned}$$

$C(t)$: concentration of the drug,
V : volume of the compartment

intravenous administration and nonlinear elimination

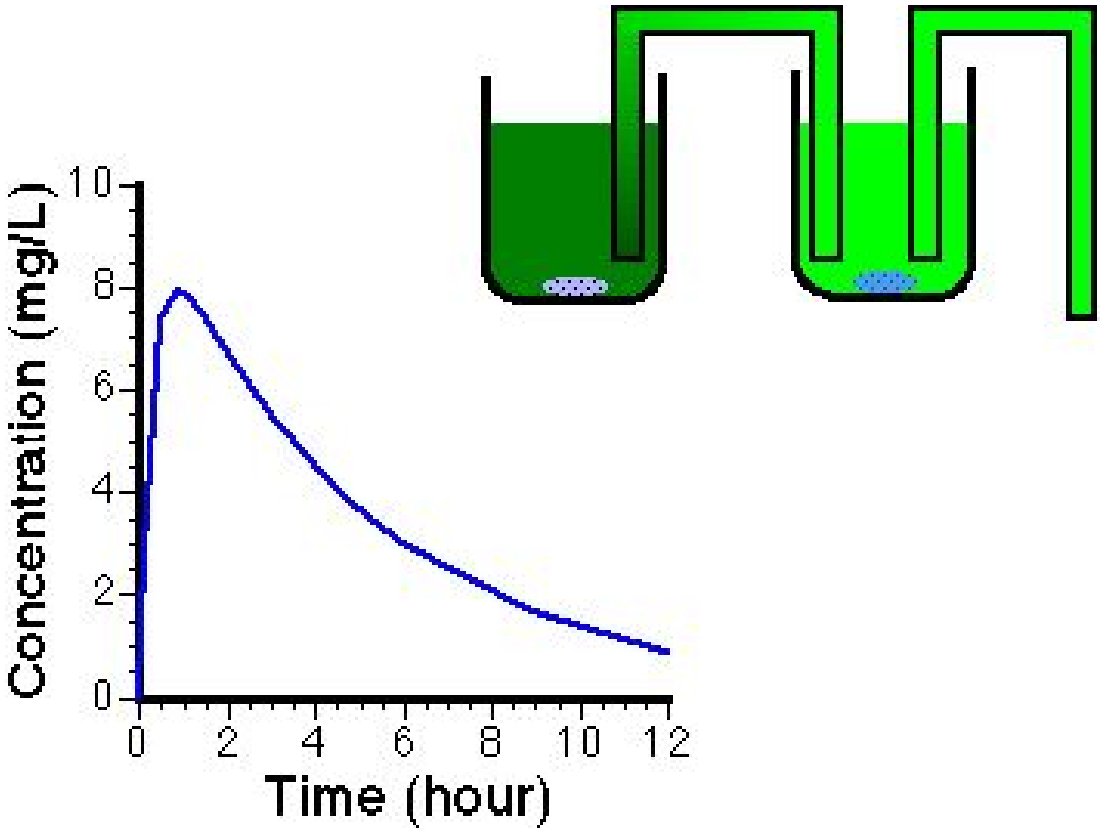
dose D (t=0) $\rightarrow$ DRUG AMOUNT $Q(t)$ $\rightarrow$ nonlinear elimination

$$\frac{dQ(t)}{dt} = -\frac{V_m\, Q(t)}{V * K_m + Q(t)}$$

$$C(t) = \frac{Q(t)}{V}$$

(V_m, K_m) : Michaelis-Menten elimination parameters,
V : volume of the compartment.

One compartment PK model
oral administration



oral administration, first-order absorbtion and elimination

dose D at time t=0

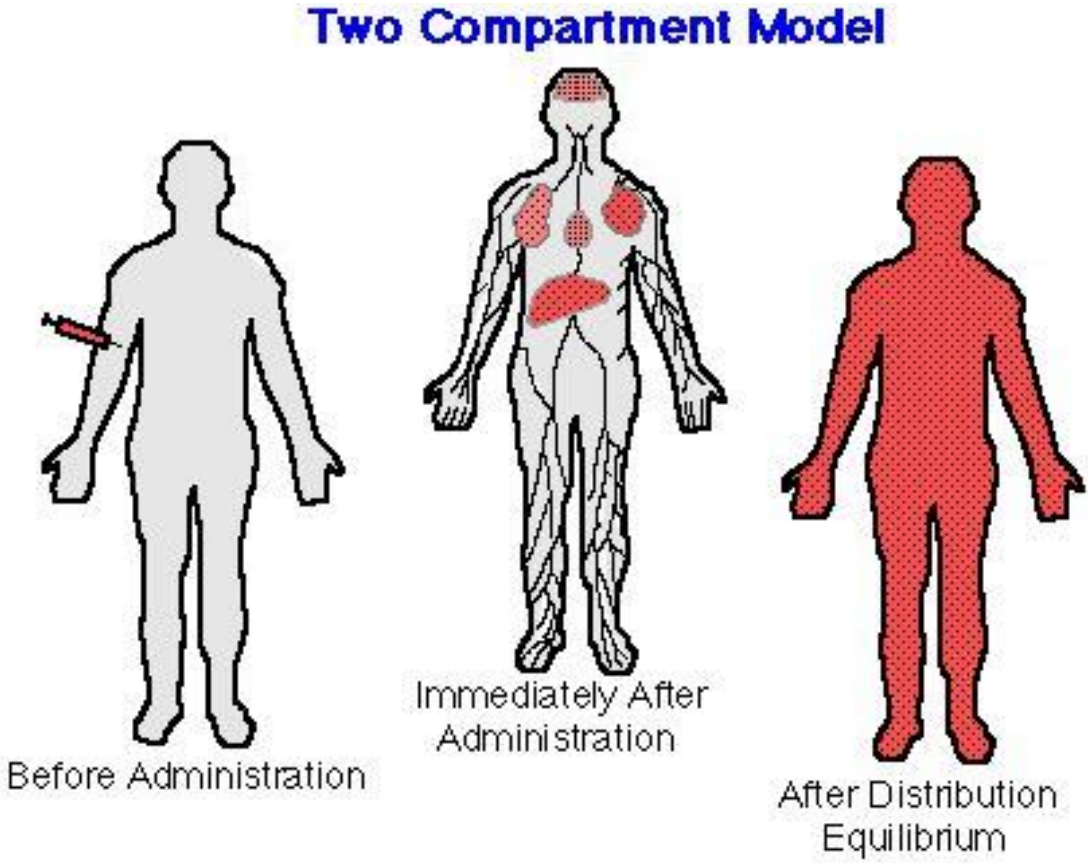
absorption (rate k_a) $\rightarrow$ DRUG AMOUNT $Q(t)$ $\rightarrow$ elimination (rate k_e)

$$\begin{aligned} \frac{dQ_a}{dt}(t) &= -k_a Q_a(t) \quad ; \quad Q_a(0) = D \\ \frac{dQ}{dt}(t) &= k_a Q_a(t) - k_e Q(t) \quad ; \quad Q(0) = 0 \end{aligned}$$

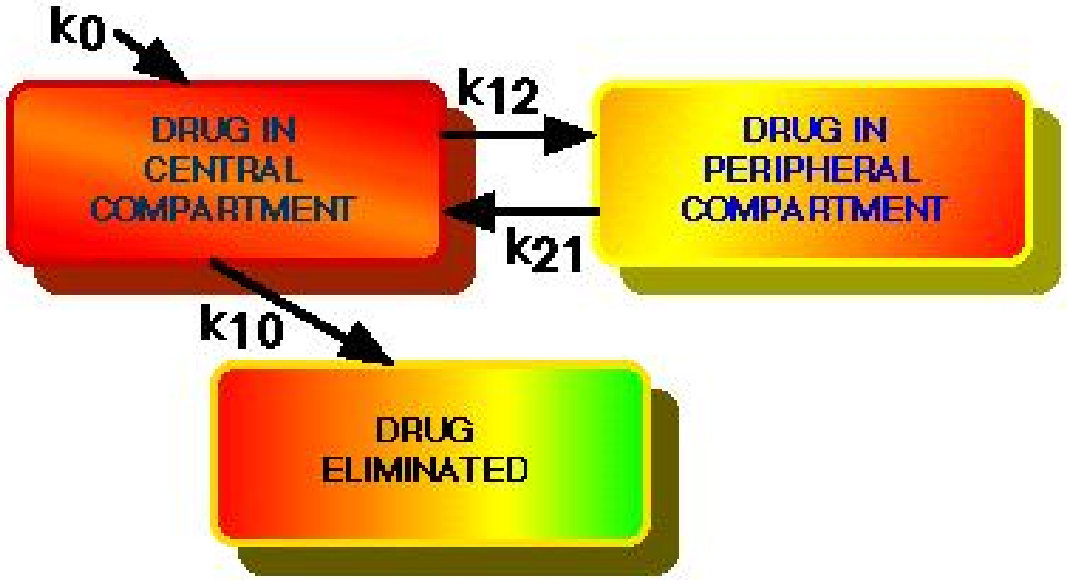
$Q_a(t)$: amount at absorption site.

$$C(t) = \frac{Q(t)}{V} = D \frac{k_a}{V(k_a - k_e)} \left(e^{-k_e t} - e^{-k_a t} \right)$$

Two compartments PK model
intravenous administration



Two compartments PK model



Two compartments PK model

$$
\begin{aligned}
\frac{dQ_a}{dt}(t) &= -k_a Q_a(t), \\
\frac{dQ_c}{dt}(t) &= k_a Q_a(t) - k_e Q_c(t) - k_{12} Q_c(t) + k_{21} Q_p(t), \\
\frac{dQ_p}{dt}(t) &= k_{12} Q_c(t) - k_{21} Q_p(t).
\end{aligned}
$$

$Q_a(t)$: amount at absorption site, $Q_a(0) = D$.

$Q_c(t)$: amount in the central compartment, $Q_c(0) = 0$.

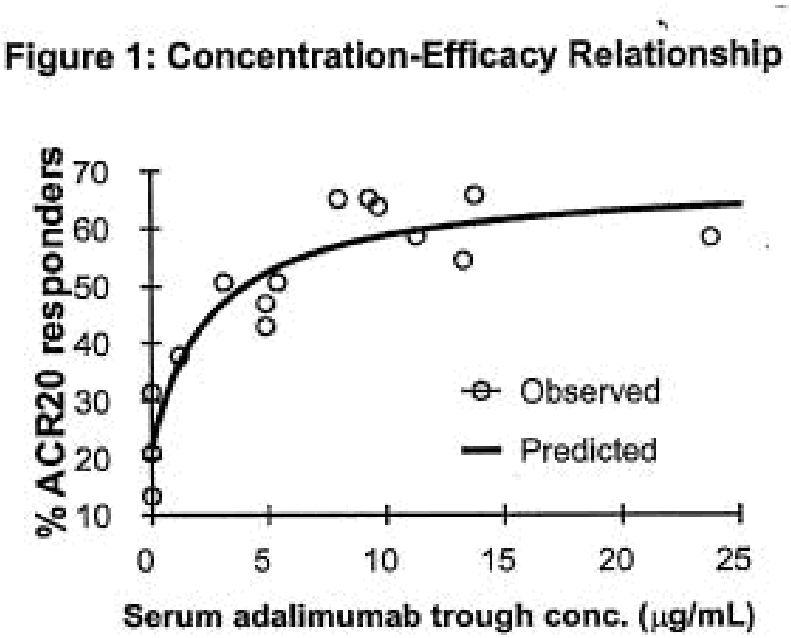
$Q_p(t)$: amount in the peripheral compartment, $Q_p(0) = 0$.

An example of pharmacodynamic model

EMax model:

$$E(t) = E_{max} \times \frac{C(t)}{C_{50} + C(t)}$$

C	E
0	0
C_{50}	$E_{max}/2$
∞	E_{max}



Outline

The regression model

$$y_j = f(x_j, \beta) + \varepsilon_j \quad , \quad 1 \leq j \leq n$$

n is the number of observations.

The regression variables, or design variables, (x_j) are **known**,

The vector of parameters β is **unknown**.

- linear model: f is a linear function of the parameters β
- non linear model: f is a non linear function of the parameters β

The regression model

$$y_j = f(x_j, \beta) + \varepsilon_j \quad , \quad 1 \leq j \leq n$$

n is the number of observations.

The regression variables, or design variables, (x_j) are **known**,

The vector of parameters β is **unknown**.

- linear model: f is a linear function of the parameters β
- non linear model: f is a non linear function of the parameters β

$$y_j = f(x_j, \beta) + \varepsilon_j$$

Here, the only random variable is the vector of residual errors $\varepsilon = (\varepsilon_j)$.

The simplest statistical model assumes that the (ε_j) are independent and identically distributed ($i.i.d$) Gaussian random variables:

$$\varepsilon_j \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

Problem: estimate the parameters of the model $\theta = (\beta, \sigma^2)$.

Maximum Likelihood Estimation

- Maximum likelihood estimation (MLE) is a popular statistical method used for fitting a statistical model to data, and providing estimates for the model's parameters.
- For a fixed set of data and underlying probability model, maximum likelihood picks the values of the model parameters that make the data "more likely" than any other values of the parameters would make them

Maximum Likelihood Estimation

Consider a family of continuous probability distributions parameterized by an unknown parameter θ, associated with a known probability density function p_θ.

Draw a vector $y = (y_1, y_2, \ldots, y_n)$ from this distribution, and then using p_θ compute the probability density associated with the observed data,

$$p_\theta(y) = p_\theta(y_1, y_2, \ldots, y_n)$$

As a function of θ with $y_1, y_2, \ldots, y_n$ fixed, this is the likelihood function

$$\mathcal{L}(\theta; y) = p_\theta(y)$$

Maximum Likelihood Estimation

Let $\theta^\star$ be the "true value" of θ.

The method of maximum likelihood estimates $\theta^\star$ by finding the value of θ that maximizes $\mathcal{L}(\theta; y)$.

This is the maximum likelihood estimator (MLE) of θ:

$$\hat{\theta} = \operatorname{Arg}\max_{\theta} \mathcal{L}(\theta; y)$$

Under certain (fairly weak) regularity conditions, the MLE is "asymptotically optimal":

- The MLE is asymptotically unbiased: $E(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is a consistant estimate of $\theta^\star$ (LLN): $\hat{\theta} \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is asymptotically normal (CLT)

$$\sqrt{n}(\hat{\theta} - \theta^\star) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1})$$

  $\mathcal{I}(\theta^\star) = -E\partial^2_\theta \log \mathcal{L}(\theta^\star; y)/n$ is the *Fisher Information Matrix*
- The MLE is asymptotically efficient, (Cramér-Rao)
  This means that no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE.

Under certain (fairly weak) regularity conditions, the MLE is "asymptotically optimal":

- The MLE is asymptotically unbiased: $E(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is a consistant estimate of $\theta^\star$ (LLN): $\hat{\theta} \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is asymptotically normal (CLT)

$$\sqrt{n}(\hat{\theta} - \theta^\star) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1})$$

  $\mathcal{I}(\theta^\star) = -E\partial^2_\theta \log \mathcal{L}(\theta^\star; y)/n$ is the *Fisher Information Matrix*
- The MLE is asymptotically efficient, (Cramér-Rao)
  This means that no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE.

Under certain (fairly weak) regularity conditions, the MLE is "asymptotically optimal":

- The MLE is asymptotically unbiased: $E(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is a consistant estimate of $\theta^\star$ (LLN): $\hat{\theta} \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is asymptotically normal (CLT)

$$\sqrt{n}(\hat{\theta} - \theta^\star) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1})$$

$\mathcal{I}(\theta^\star) = -E\partial^2_\theta \log \mathcal{L}(\theta^\star; y)/n$ is the *Fisher Information Matrix*

- The MLE is asymptotically efficient, (Cramér-Rao)
  This means that no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE.

Under certain (fairly weak) regularity conditions, the MLE is "asymptotically optimal":

- The MLE is asymptotically unbiased: $E(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is a consistant estimate of $\theta^\star$ (LLN): $\hat{\theta} \underset{n\to\infty}{\longrightarrow} \theta^\star$
- The MLE is asymptotically normal (CLT)

$$\sqrt{n}(\hat{\theta} - \theta^\star) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1})$$

$\mathcal{I}(\theta^\star) = -E\partial_\theta^2 \log \mathcal{L}(\theta^\star; y)/n$ is the *Fisher Information Matrix*

- The MLE is asymptotically efficient, (Cramér-Rao)
  This means that no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE.

$$\begin{array}{rcl} y_j & = & f(x_j, \beta) + \varepsilon_j \quad , \quad 1 \leq j \leq n \\ \varepsilon_j & \sim_{i.i.d.} & \mathcal{N}(0, \sigma^2) \end{array}$$

$$y \sim \mathcal{N}(f(x_j, \beta), \sigma^2 I_n)$$

$$\mathcal{L}(\theta; y) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - f(x_j,\beta))^2}$$

$$\hat{\beta} = \operatorname{Arg}\max_{\beta} \mathcal{L}(\beta; y) = \operatorname{Arg}\min_{\beta} \sum_{j=1}^{n} (y_j - f(x_j, \beta))^2$$

(Maximum Likelihood estimate of β = Least-Square estimate of β)

The linear regression model

$$
\begin{aligned}
y_1 &= x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1p}\beta_p + \varepsilon_1 \\
y_2 &= x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2p}\beta_p + \varepsilon_2 \\
&\vdots \\
y_n &= x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots + x_{np}\beta_p + \varepsilon_n
\end{aligned}
$$

$$
\boxed{Y = X\,\beta + \varepsilon}
$$

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma^2) \\ \theta &= (\beta, \sigma^2) \end{aligned}$$

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

$$\mathcal{L}(\theta; y) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\|y - X\beta\|^2}$$

$$\hat{\beta} = \operatorname{Arg}\max_{\beta} \mathcal{L}(\beta; y) = \operatorname{Arg}\min_{\beta} \|y - X\beta\|^2$$

$$y = X\beta + \varepsilon$$

$$\begin{aligned}
\hat{\beta} &= \operatorname{Arg\,min}_{\beta} \|y - X\beta\|^2 \\
&= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + \varepsilon) \\
&= \beta + (X'X)^{-1}X'\varepsilon \\
E(\hat{\beta}) &= \beta \\
Var(\hat{\beta}) &= \sigma^2(X'X)^{-1}
\end{aligned}$$

$$-\log \mathcal{L}(\theta^\star; y) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|y - X\beta\|^2$$

$$\mathcal{I}(\beta) = -\frac{1}{n}E\partial^2_\beta \log \mathcal{L}(\beta, \sigma^2; y) = \frac{1}{n\sigma^2}(X'X)$$

$$y = X\beta + \varepsilon$$

$$\begin{aligned}
\hat{\beta} &= \operatorname{Arg\,min}_{\beta} \|y - X\beta\|^2 \\
&= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + \varepsilon) \\
&= \beta + (X'X)^{-1}X'\varepsilon \\
E(\hat{\beta}) &= \beta \\
Var(\hat{\beta}) &= \sigma^2(X'X)^{-1}
\end{aligned}$$

$$\begin{aligned}
-\log \mathcal{L}(\theta^\star; y) &= \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|y - X\beta\|^2 \\
\mathcal{I}(\beta) &= -\frac{1}{n}E\partial^2_\beta \log \mathcal{L}(\beta, \sigma^2; y) = \frac{1}{n\sigma^2}(X'X)
\end{aligned}$$

$$y = X\beta + \varepsilon$$

Let $V = Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ be the variance-covariance matrix of $\hat{\beta}$.

The diagonal elements of V are the variances of the components of $\hat{\beta}$:

- $V_{k,k}$ is the variance of $\hat{\beta}_k$
- $\sqrt{V_{k,k}}$ is the *standard error* (*s.e.*) of $\hat{\beta}_k$
- 90% confidence interval for β_k:

$$[\hat{\beta}_k - 1.645\sqrt{V_{k,k}}; \hat{\beta}_k + 1.645\sqrt{V_{k,k}}]$$

Optimal design

- Optimal designs are a class of experimental designs that are optimal with respect to some statistical criterion.
- In the design of experiments for estimating statistical models, optimal designs allow parameters to be estimated without bias and with minimum-variance.
- A non-optimal design requires a greater number of experimental runs to estimate the parameters with the same precision as an optimal design.
- In practical terms, optimal experiments can reduce the costs of experimentation.
- Fisher information is widely used in optimal experimental design. Because of the reciprocity of estimator-variance and Fisher information, minimizing the variance corresponds to maximizing the information.
- D-optimal design maximizes the determinant of the Fisher information matrix $\mathcal{I}(\theta^\star)$.

A D-optimal design maximizes the determinant of $X'X$.

Example:

$$y_j = a + bx_j + e_j$$

Here,

$$\begin{aligned} \frac{1}{n}|X'X| &= \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2 \\ &= \text{Variance of } (x_1, x_2, \ldots, x_n) \end{aligned}$$

For a fixed number of measurements n, an optimal design has maximum variance.

Model selection

1. **Compute the likelihood of the different models**
   - Let $\hat{\theta}_{\mathcal{M}}$ be the maximum likelihood estimate of θ for model $\mathcal{M}$:

$$\hat{\theta}_{\mathcal{M}} = \text{Arg}\max_{\theta} \mathcal{L}_{\mathcal{M}}(\theta; y)$$

   - Let $\mathcal{L}_{\mathcal{M}} = \mathcal{L}_{\mathcal{M}}(\hat{\theta}_{\mathcal{M}}; y)$ be the likelihood of model $\mathcal{M}$.

   Selecting the "most likely" models by comparing the likelihoods favor models of high dimension (with many parameters)!

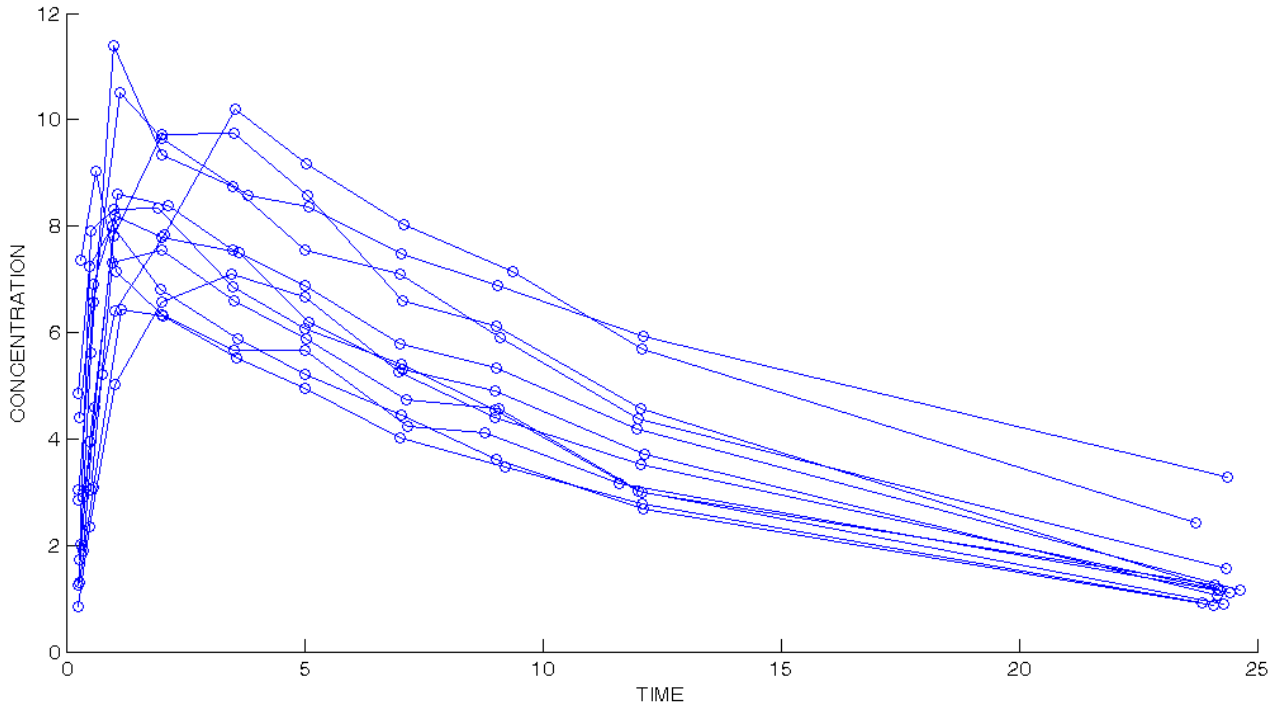
2. **Penalize the models of high dimension**
   Select the model $\hat{\mathcal{M}}$ that minimizes the penalized criteria

$$-2\mathcal{L}_{\mathcal{M}} + pen(\mathcal{M})$$

   Bayesian Information Criteria (BIC) : $pen(\mathcal{M}) = \log(n) \times \dim(\mathcal{M})$.
   Akaike Information Criteria (AIC) : $pen(\mathcal{M}) = 2\dim \times (\mathcal{M})$.

Outline

12 patients:



Each individual curve is described by the same parametric model, with its own individual parameters.

- inter-subject variation in concentrations for same dose
- inter-subject variation in response for same dose

$\Longrightarrow$ each subject may have same model but with different PK/PD parameters.

Complications:

- times of measurements depend on the subject
- observations contain errors (measurement, model misspecification,...)
- observations above some limit of quantification (concentration, viral load,...)
- part of the inter-variability explained by some known covariates (weight, age, gender,...)
- ...

The mixed effects model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

$y_{ij} \in \mathbb{R}$ is the jth observation of subject i,

N is the number of subjects

n_i is the number of observations of subject i.

The regression variables, or design variables, (x_{ij}) are **known**,

The individual parameters (ψ_i) are **unknown**.

The mixed effects model

$$\begin{aligned} y_{ij} &= f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \\ \psi_i &= h(C_i, \beta, \eta_i) \end{aligned}$$

C_i is a vector of covariates
β is a $p-$vector of fixed effects
η_i is a $q-$vector of random effects

$$\begin{aligned} \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \\ \eta_i &\sim \mathcal{N}(0, \Omega) \end{aligned}$$

Ω is the $q \times q$ variance-covariance matrix of the random effects

(Hyper)parameters of the model: $\theta = (\beta, \Omega, \sigma^2)$.

The mixed effects model

$$\begin{aligned} y_{ij} &= f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \\ \psi_i &= h(C_i, \beta, \eta_i) \end{aligned}$$

C_i is a vector of covariates
β is a $p-$vector of fixed effects
η_i is a $q-$vector of random effects

$$\begin{aligned} \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \\ \eta_i &\sim \mathcal{N}(0, \Omega) \end{aligned}$$

Ω is the $q \times q$ variance-covariance matrix of the random effects

(Hyper)parameters of the model: $\theta = (\beta, \Omega, \sigma^2)$.

Estimation

- Estimate the set of population parameters θ,
- Compute confidence intervals,

Model selection

- Determine if a parameter varies in the population
- Select the best combination of covariates
- Compare several treatments
- ...

Optimal design

- Determine the design (the measurement times) that yields the most accurate estimation of the model

The maximum likelihood estimator of $\theta = (\beta, \Omega, \sigma^2)$ maximizes

$$
\begin{aligned}
\mathcal{L}(\theta; y) &= \prod_{i=1}^{N} \mathcal{L}_i(\theta; y_i) \\
\mathcal{L}_i(\theta; y_i) &= \int p(y_i, \eta_i; \theta) d\eta_i \\
&= \int p(y_i|\eta_i; \theta) p(\eta_i; \theta) d\eta_i \\
&= C \int \sigma^{-n_i} |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\|y_i - f(x_i; h(Ci, \beta, \eta_i))\|^2 - \frac{1}{2}\eta_i' \Omega^{-1} \eta_i} d\eta_i
\end{aligned}
$$

Example: $\psi_i = \beta + \eta_i$

$$
\mathcal{L}_i(\theta; y_i) = C \int \sigma^{-n_i} |\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\|y_i - f(x_i; \psi_i))\|^2 - \frac{1}{2}(\psi_i - \beta)' \Omega^{-1} (\psi_i - \beta)} d\psi_i
$$

Assume that $\theta = (\beta, \Omega, \sigma^2)$ is known (or was estimated previously)
$\hat{\psi}_i$ maximizes the conditional distribution $p(\psi_i|y_i;\theta)$:

$$\begin{aligned} p(\psi_i|y_i;\theta) &= \frac{p(\psi_i, y_i;\theta)}{p(y_i;\theta)} \\ &= \frac{p(y_i|\psi_i;\theta)p(\psi_i;\theta)}{p(y_i;\theta)} \\ &= C\ p(y_i|\psi_i;\theta)p(\psi_i;\theta) \end{aligned}$$

Example: $\psi_i = \beta + \eta_i$

$$p(\psi_i|y_i;\theta) = e^{-\frac{1}{2\sigma^2}\|y_i - f(x_i;\psi_i))\|^2 - \frac{1}{2}(\psi_i-\beta)'\Omega^{-1}(\psi_i-\beta)}$$

Then, $\hat{\psi}_i$ minimizes a penalized least-square criteria:

$$\hat{\psi}_i = \mathrm{Arg}\min_{\psi}\left\{\|y_i - f(x_i;\psi))\|^2 + \sigma^2(\psi-\beta)'\Omega^{-1}(\psi-\beta)\right\}$$

Some existing methods

1. Methods based on individual estimates

   i) Estimate the individual parameters (ψ_i),

   ii) Estimate θ using $(\widehat{\psi}_i)$.

$\Longrightarrow$ Requires a large number of observations per subject.

Some existing methods

2. Methods based on approximations of the likelihood

- First order methods (FO, Beal and Sheiner, 1982)

$$
\begin{aligned}
y_{ij} &= f(x_{ij}, \psi_i) + \varepsilon_{ij} \\
\psi_i &= \beta + \eta_i \\
y_{ij} &\approx f(x_{ij}, \beta) + \frac{\partial f}{\partial \psi_i}(x_{ij}, \beta)\eta_i + \varepsilon_{ij}
\end{aligned}
$$

i) NONMEM package (very popular in pharmacokinetics)

ii) SAS proc NLMIXED (using the method=firo option)

Some existing methods

2. Methods based on approximations of the likelihood

- First order conditional methods (FOCE, Lindstrom and Bates, 1990)

$$y_{ij} \approx f(x_{ij}, \hat{\psi}_i) + \frac{\partial f}{\partial \psi_i}(x_{ij}, \hat{\psi}_i)(\psi_i - \hat{\psi}_i) + \varepsilon_{ij}$$

$\hat{\psi}_i$ maximizes the conditional distribution $p(\psi_i | y_i; \theta)$

i) NONMEM package (FOCE option)

ii) SAS proc NLMIXED (using the method=eblup option)

iii) Splus/R function NLME

- theoretical drawbacks: no well-known statistical properties of the algorithm,
- practical drawbacks: very sensitive to the initial guess, does not always converge, poor estimation of some parameters,. . .

Some existing methods

3. Methods based on numerical approximations of the likelihood

- Laplace method
- Gaussian quadrature method

- nice theoretical properties: maximum likelihood estimation is performed,
- practical drawbacks : limited to few random effects.

The MONOLIX Group

MOdèles NOn LInéaires à effets miXtes

This multi-disciplinary group, born in october 2003 develop activities in the field of mixed effect models. It involves scientists with varied backgrounds, interested both in the study and applications of these models:

- academic statisticians from several universities of Paris (theoretical developments),
- researchers from INSERM (U738, applications in pharmacology)
- researchers from INRA (applications in agronomy, animal genetics...),
- scientists from the medical faculty of Lyon-Sud University (applications in oncology).

The MONOLIX Group

The objectives of the group are multiple:

- develop new methodologies,
- study the theoretical properties of these methodologies,
- apply these methodologies to realistic problems,
- implement these methodologies in a free software, available to the whole community.

Outline

MONOLIX 2
an "academic" (but promising) software

MONOLIX 2 is an open-source software using Matlab (a StandAlone version is available)

MONOLIX 2 was developed from April 2007 to October 2008 :

- Version 2.1: April 2007
- Version 2.2: June 2007
- Version 2.3: November 2007
  - release 2.3.1: March 2008 (C++ package)
  - release 2.3.2: April 2008 (categorical covariates ; transformation of the individual parameters ...)
  - release 2.3.4: May 2008 (Inter Occasion Variability)
- Version 2.4: July 2008 (3 cpts PK models ; effect compartment models ; "NMTRAN-like" interpreter)
  - release 2.4 stable: October 2008

About 100 downloads per month

Academics

Universities : Iowa, Utah, Massachusetts, Kentucky, Maryland, Pennsylvania, Pittsburgh, Buffalo, Brown, Uppsala, Utrecht, Bern, Gdansk, Belfast, Melbourne, Auckland, Cape Town, Teheran, Karachi, Heilongjiang, Kyushu, Kyoto, Yogyaka, Naresuan, Okayama, Buenos-Aires,...
INSERM, CHU, CNRS, INRA, ENVT,...

Industry

Novartis, Roche, Johnson & Johnson, Sanofi-Aventis, Pfizer, GSK, Merck, BMS, UCB, Servier, Otsuka, Tibotec, Solvay, Abbott, Amgen, Chugai, Merrimack, Novo Nordisk,...

Consulting companies

Exprimo, Pharsight, Nektar, Freise, Rosa, ...

- PAGE 2009, St-Petersbourg, Russie, Juin 2009
- Université de Buffalo, USA, Mars 2009
- Université de Sheffield, Angleterre, Janvier 2009
- Hoffmann-La Roche, Bâle, Suisse, Décembre 2008
- PAGE 2008, Marseille, France, Juin 2008
- Johnson & Johnson, Beerse, Belgique, Mai 2008
- Novartis Pharma, Cambridge, USA, Mai 2008
- Novartis Pharma, East Hanover, USA, Mai 2008
- UCB, Braine l'Allaud, Belgique, Mars 2008
- Novartis Pharma, Bâle, Suisse, Novembre 2007
- PAGANZ 2007, Singapour, Février 2007

MONOLIX 3
toward a "professional" software

- The MONOLIX project consists primarily in developing the next versions of the MONOLIX software with a view to raising its level of functionalities and responding to major requirements of the bio-pharmaceutical industry.
- The MONOLIX project is a 3-year software development project by a 5-engineer Monolix team.
- The MONOLIX Project is carried out by INRIA, and sponsored by the Industry
- Members of the project: Novartis, Roche, Johnson & Johnson, Sanofi-Aventis,
- The MONOLIX Scientific Guidance Committee involves representatives of the sponsors.

- Version 3.1: July 2009 (discrete data models, HMM, complex PK models, ...)

The algorithms used in MONOLIX

Intensive use of powerful and well-known algorithms in the MONOLIX software:

- **Estimation of the population parameters:** Maximum likelihood estimation with the SAEM (Stochastic Approximation of EM) algorithm, combined with MCMC (Markov Chain Monte Carlo) and Simulated Annealing,
- **Estimation of the individual parameters:** Estimation/Maximization of the conditional distributions with MCMC,
- **Estimation of the objective (likelihood) function:** Monte Carlo and minimum variance Importance Sampling,
- **Model selection and assessment:** Information criteria (AIC, BIC), Statistical Tests (LRT, Wald test), Goodness of fit plots (Individual fits, Weighted residuals, NPDE, VPC,...).

1) Some examples without covariates

$$\begin{aligned} \psi_i &= (\psi_{ik}, 1 \leq k \leq p) \\ &= h(\beta, \eta_i) \\ \psi^{pop} &= h(\beta, 0) \quad \text{(population parameter)} \end{aligned}$$

- Normal distribution (ψ_{ik} can take any value in $\mathbb{R}$)

$$\psi_{ik} = \beta_k + \eta_{ik} = \psi_k^{pop} + \eta_{ik}$$

- log-normal distribution (assuming that $\psi_{ik} > 0$)

$$\psi_{ik} = e^{\beta_k + \eta_{ik}} = \psi_k^{pop} e^{\eta_{ik}}$$

- logit transformation (assuming that $0 < \psi_{ik} < 1$)

$$\psi_{ik} = \frac{1}{1 + e^{-\beta_k - \eta_{ik}}} = \frac{\psi_k^{pop}}{\psi_k^{pop} + (1 - \psi_k^{pop}) e^{-\eta_{ik}}}$$

2) An example with Weight as a covariate

$$\begin{aligned} \psi_i &= (\psi_{ik}, 1 \leq k \leq p) \\ &= h(\mathrm{W}_i, \beta, \eta_i) \end{aligned}$$

$$\begin{aligned} \psi_{ik} &= e^{\beta_{k1}+\eta_{ik}} \left(\frac{\mathrm{W}_i}{\mathrm{W}^{\mathrm{pop}}}\right)^{\beta_{k2}} \\ &= \psi_k^{pop} \left(\frac{\mathrm{W}_i}{\mathrm{W}^{\mathrm{pop}}}\right)^{\beta_{k2}} e^{\eta_{ik}} \\ \log(\psi_{ik}) &= \log(\psi_k^{pop}) + \beta_{k2} \log\left(\frac{\mathrm{W}_i}{\mathrm{W}^{\mathrm{pop}}}\right) + \eta_{ik} \end{aligned}$$

Assume that some categorical covariate C_i takes M values

$$\psi_{ik} = \psi_k^{ref} + \sum_{m=1}^{M} \beta_{k,m} \mathbb{1}_{C_i=m} + \eta_{ik}$$

- $m^\star$: reference group $\iff$ $\beta_{k,m^\star} = 0$
- The variances of the random effects can also depend on this categorical covariate.

Assume that some categorical covariate C_i takes M values

$$\psi_{ik} = \psi_k^{ref} + \sum_{m=1}^{M} \beta_{k,m} \mathbb{1}_{C_i=m} + \eta_{ik}$$

- $m^\star$: reference group $\iff$ $\beta_{k,m^\star} = 0$
- The variances of the random effects can also depend on this categorical covariate.

$$y_{ij} = f(x_{ij}, \psi_i) + g(x_{ij}, \psi_i)\varepsilon_{ij} \quad , \quad 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

The residual errors (ε_{ij}) are supposed to be *i.i.d.* Gaussian random variables with mean zero and variance $\sigma^2 = 1$.

Example : $g(x_{ij}, \psi_i) = a + bf^c(x_{ij}, \psi_i)$

- *the constant error model:* $y = f + a\varepsilon$,
- *the proportional error model:* $y = f(1 + b\varepsilon)$,
- *combined error model:* $y = f + (a + bf)\varepsilon$.

Extension:

$$t(y_{ij}) = t(f(x_{ij}, \psi_i)) + g(x_{ij}, \psi_i)\varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

Examples

- *the exponential error model:* $t(y) = \log(y)$:

$$y = fe^{g\varepsilon}$$

,

- *the logit error model:* $t(y) = \log(y/(1-y))$,

$$y = \frac{f}{f + (1-f)e^{-g\varepsilon}}$$

- Statistical model:

$$\begin{aligned} y_{ij} &= f(x_{ij},\psi_i) + g(x_{ij},\psi_i)\varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \\ \varepsilon_{ij} &\sim \mathcal{N}(0,1) \end{aligned}$$

- Observed data

$$y_{ij}^{obs} = \begin{cases} y_{ij} & \text{if} \quad y_{ij} > LOQ \\ LOQ & \text{if} \quad y_{ij} \leq LOQ \end{cases}$$

- Left-censored data $y^{cens} = \{y_{ij} | y_{ij} \leq LOQ\}$

$$\begin{aligned} y_{ij}^{(1)} &= f_1(x_{ij}^{(1)}, \psi_i) + \varepsilon_{ij}^{(1)} \\ &\vdots \\ y_{ij}^{(L)} &= f_L(x_{ij}^{(L)}, \psi_i) + \varepsilon_{ij}^{(L)} \end{aligned}$$

PKPD model: the input of the PD model $x_{ij}^{(2)}$ is the output of the PK model $f_1(x_{ij}^{(1)}, \psi_i)$ (the concentration).

Viral dynamics: $y^{(1)}$ is the viral load and $y^{(2)}$ is the $CD4+$ count.

$$y_{ikj} = f(x_{ikj}, \psi_{ik}) + g(x_{ikj}, \psi_{ik})\varepsilon_{ikj}$$

- $i = 1, \ldots, N$ is the subject
- $k = 1, \ldots, K$ is the occasion
- $j = 1, \ldots, n_{ik}$ is the measure
- y_{ikj} is the jth observation of occasion k and subject i
- ψ_{ik} individual parameter of subject i at occasion k

The non-linear mixed effects model

Modeling the inter-occasion variability

$$\psi_{ik} = \psi^{pop} + \eta_i + \kappa_{ik}$$

- μ (px1) population parameter
- η_i (px1) random effect of subject i: $\eta_i \sim \mathcal{N}(0, \Omega)$
- κ_{ik} (px1) random effect of subject i at occasion k: $\kappa_{ik} \sim \mathcal{N}(0, \Gamma)$
- η_i et κ_{ik} are assumed to be independent
- Ω (pxp) inter-subject variability
- Γ (pxp) inter-occasion variability

Model with covariate:

$$\psi_{ik} = \mu + \beta C_i + \tilde{\beta}\tilde{C}_{ik} + \eta_i + \kappa_{ik}$$

The non-linear mixed effects model

Modeling the inter-occasion variability

$$\boxed{\psi_{ik} = \psi^{pop} + \eta_i + \kappa_{ik}}$$

- μ (px1) population parameter
- η_i (px1) random effect of subject i: $\eta_i \sim \mathcal{N}(0, \Omega)$
- κ_{ik} (px1) random effect of subject i at occasion k: $\kappa_{ik} \sim \mathcal{N}(0, \Gamma)$
- η_i et κ_{ik} are assumed to be independent
- Ω (pxp) inter-subject variability
- Γ (pxp) inter-occasion variability

Model with covariate:

$$\boxed{\psi_{ik} = \mu + \beta C_i + \tilde{\beta}\tilde{C}_{ik} + \eta_i + \kappa_{ik}}$$

Outline

The incomplete data model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

We are in a classical framework of *"incomplete data"*:

- the measurement $y = (y_{ij}, 1 \leq i \leq N \, , 1 \leq j \leq n_i)$ are the *"observed data"*
- the individual random parameters $\phi = (\psi_i, \, 1 \leq i \leq N)$, are the *"non observed data"*,
- the *"complete data"* of the model is (y, ϕ).

Estimation of the population parameters:
compute $\widehat{\theta}$, the maximum likelihood estimate of the unknown set of parameters $\theta = (\beta, \Omega, \sigma^2)$, by maximizing the likelihood of the observations $\ell(y, \theta)$, without any approximation on the model.

The incomplete data model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

We are in a classical framework of *"incomplete data"*:

- the measurement $y = (y_{ij}, 1 \leq i \leq N \, , 1 \leq j \leq n_i)$ are the *"observed data"*
- the individual random parameters $\phi = (\psi_i, \, 1 \leq i \leq N)$, are the *"non observed data"*,
- the *"complete data"* of the model is (y, ϕ).

Estimation of the population parameters:
compute $\widehat{\theta}$, the maximum likelihood estimate of the unknown set of parameters $\theta = (\beta, \Omega, \sigma^2)$, by maximizing the likelihood of the observations $\ell(y, \theta)$, without any approximation on the model.

The incomplete data model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

We are in a classical framework of *"incomplete data"*:

- the measurement $y = (y_{ij}, 1 \leq i \leq N \, , 1 \leq j \leq n_i)$ are the *"observed data"*
- the individual random parameters $\phi = (\psi_i, \, 1 \leq i \leq N)$, are the *"non observed data"*,
- the *"complete data"* of the model is (y, ϕ).

Estimation of the individual parameters:
compute/maximize the conditional distributions of the individual parameters $p(\phi_i | y_i; \widehat{\theta} \,)$, without any approximation on the model.

The incomplete data model

$$y_{ij} = f(x_{ij}, \psi_i) + \varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i$$

We are in a classical framework of *"incomplete data"*:

- the measurement $y = (y_{ij}, 1 \leq i \leq N \, , 1 \leq j \leq n_i)$ are the *"observed data"*
- the individual random parameters $\phi = (\psi_i, \, 1 \leq i \leq N)$, are the *"non observed data"*,
- the *"complete data"* of the model is (y, ϕ).

Estimation of the likelihood function:
compute the observed likelihood $\ell(y, \widehat{\theta}\,)$, without any approximation on the model.

The EM algorithm (Expectation-Maximization)
(Dempster, Laird et Rubin, JRSSB, 1977)

Since ϕ is not observed, $\log p(y, \phi; \theta)$ cannot be directly used for estimating θ. Then

Iteration k of the algorithm:

- step E : evaluate the quantity

$$Q_k(\theta) = \mathbb{E}[\log p(y, \phi; \theta) | y; \theta_{k-1}]$$

- step M : update the estimation of θ:

$$\theta_k = Argmax \ \ Q_k(\theta)$$

The EM algorithm (Expectation-Maximization)
(Dempster, Laird et Rubin, JRSSB, 1977)

Since ϕ is not observed, $\log p(y, \phi; \theta)$ cannot be directly used for estimating θ. Then

Iteration k of the algorithm:

- step E : evaluate the quantity

$$\boxed{Q_k(\theta) = \mathbb{E}[\log p(y, \phi; \theta) | y; \theta_{k-1}]}$$

- step M : update the estimation of θ:

$$\boxed{\theta_k = Argmax \;\; Q_k(\theta)}$$

The EM algorithm (Expectation-Maximization)
(Dempster, Laird et Rubin, JRSSB, 1977)

Since ϕ is not observed, $\log p(y, \phi; \theta)$ cannot be directly used for estimating θ. Then

Iteration k of the algorithm:

- step E : evaluate the quantity

$$\boxed{Q_k(\theta) = \mathbb{E}[\log p(y, \phi; \theta) | y; \theta_{k-1}]}$$

- step M : update the estimation of θ:

$$\boxed{\theta_k = Argmax \;\; Q_k(\theta)}$$

Convergence of EM

Theorem

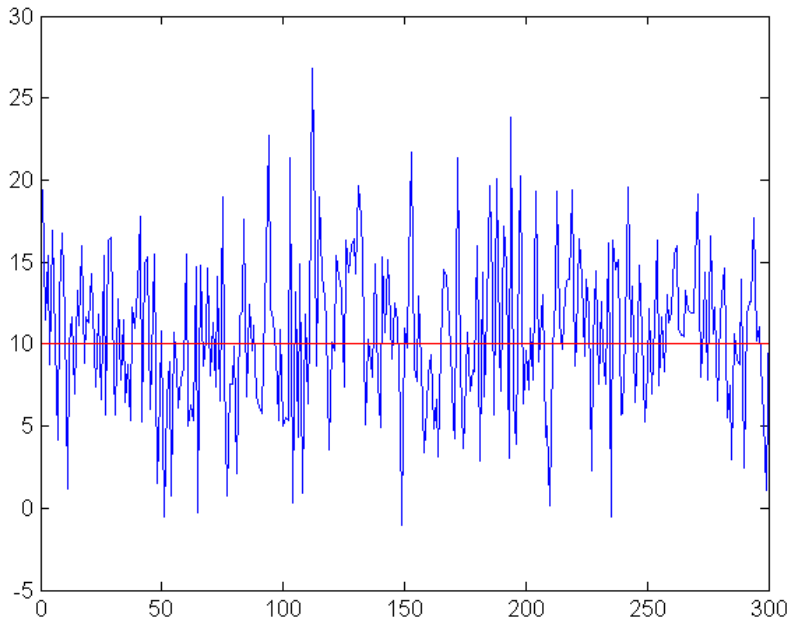
Convergence of (θ_k) to a stationary point $\widehat{\theta}_\ell$ of the observed likelihood is ensured under some regularity conditions.

Some practical drawbacks of EM:

- Convergence depends on the initial guess.
- Slow convergence of EM.
- Evaluation of $Q_k(\theta) = E[\log p(y, \phi; \theta)|y; \theta_{k-1}]$ during step E.

Assume that we can observe (or we can draw) $x_1, x_2, \ldots$

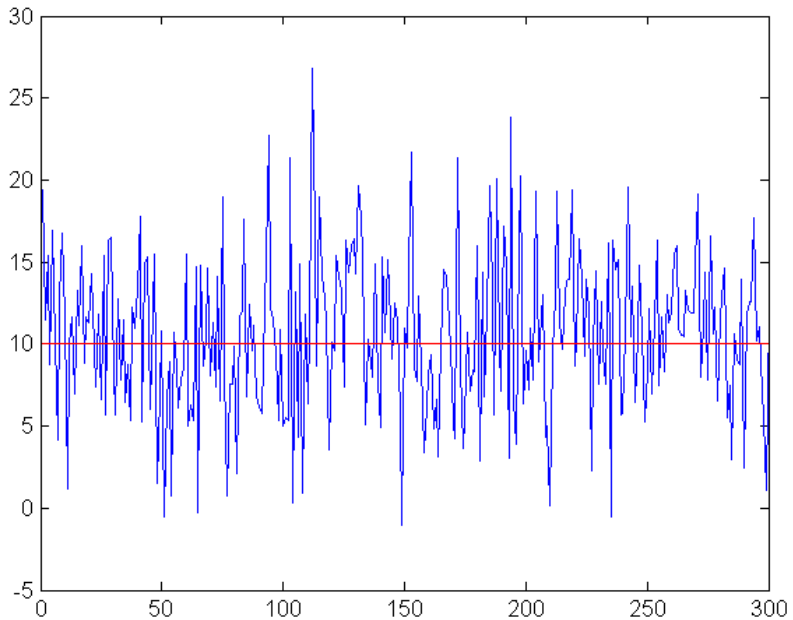
$$E(x_k) = m \quad ; \quad Var(x_k) = \sigma^2$$



We aim to estimate $m = E(x_k)$

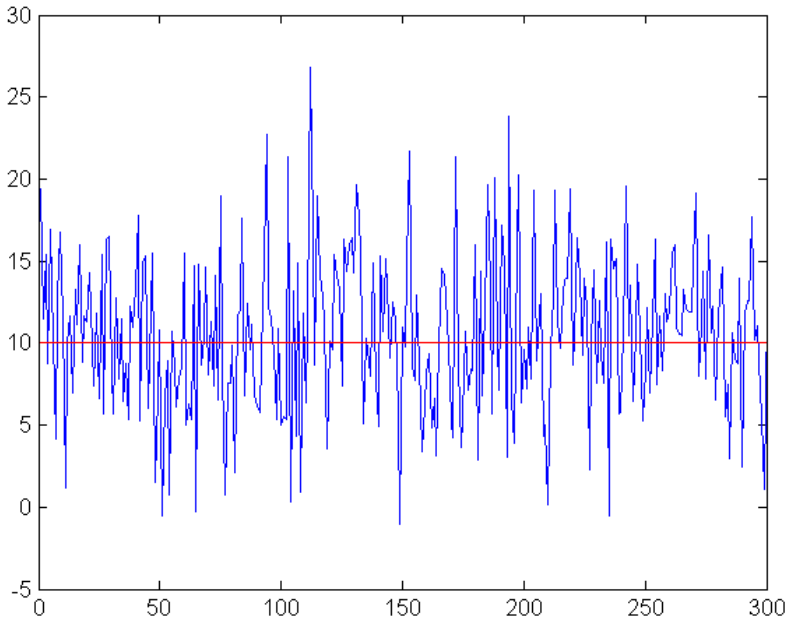
Assume that we can observe (or we can draw) $x_1, x_2, \ldots$

$$E(x_k) = m \quad ; \quad Var(x_k) = \sigma^2$$



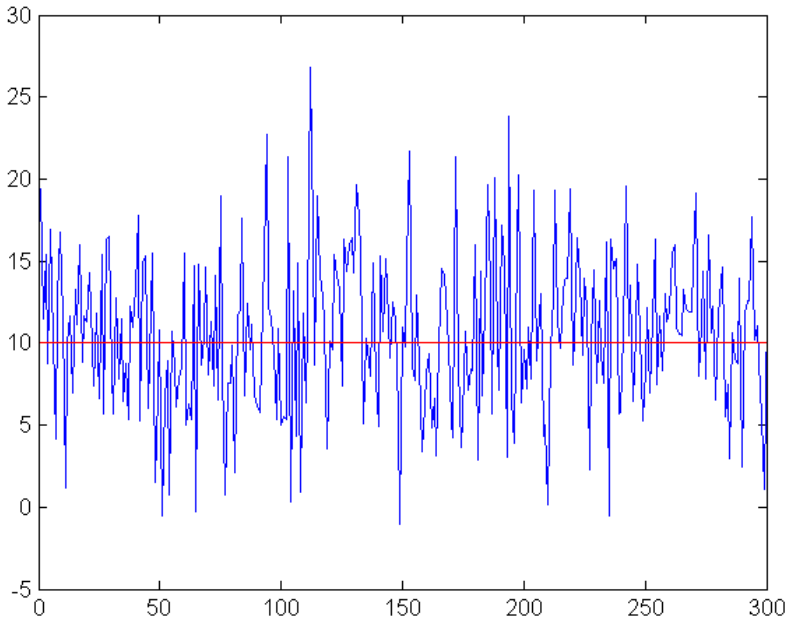
We aim to estimate $m = E(x_k)$

1) approximate m by x_k



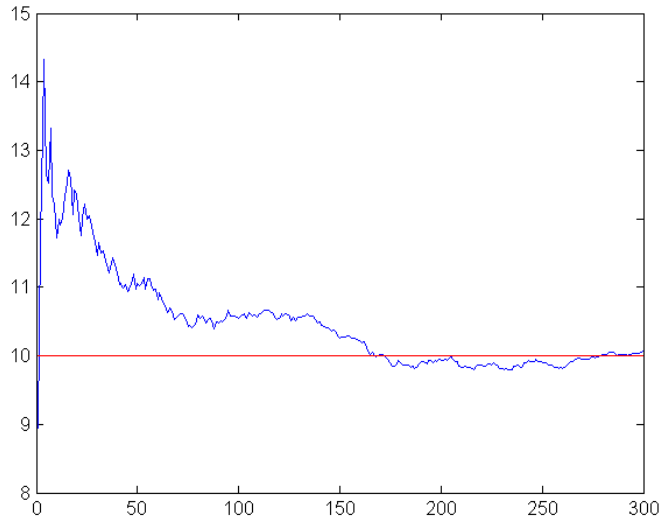
- unbiased estimate since $E(x_k) = m$
- not consistant estimate since $Var(x_k) = \sigma^2$.

1) approximate m by x_k



- unbiased estimate since $E(x_k) = m$
- not consistant estimate since $Var(x_k) = \sigma^2$.

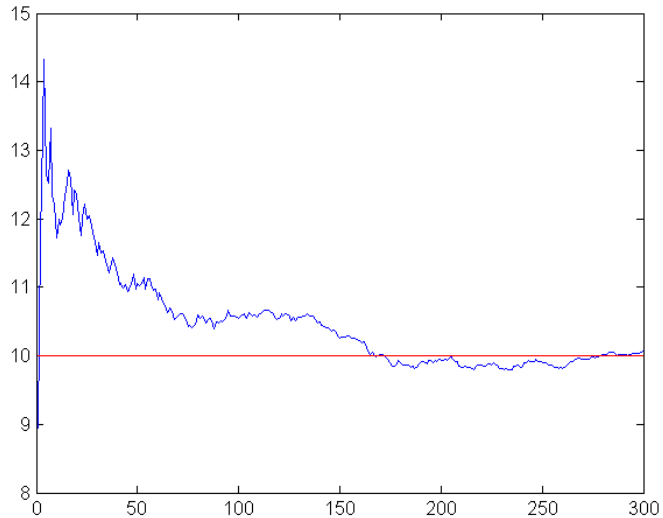
2) approximate m by $\overline{x}_k = 1/k \sum_{i=1}^{k} x_i$



- unbiased estimate since $E(\overline{x}_k) = m$,
- consistant estimate since $Var(\overline{x}_k) \to 0$.
- Thus, $\overline{x}_k \to m$ when $k \to \infty$.

$$\overline{x}_k = \overline{x}_{k-1} + \frac{1}{k}(x_k - \overline{x}_{k-1})$$

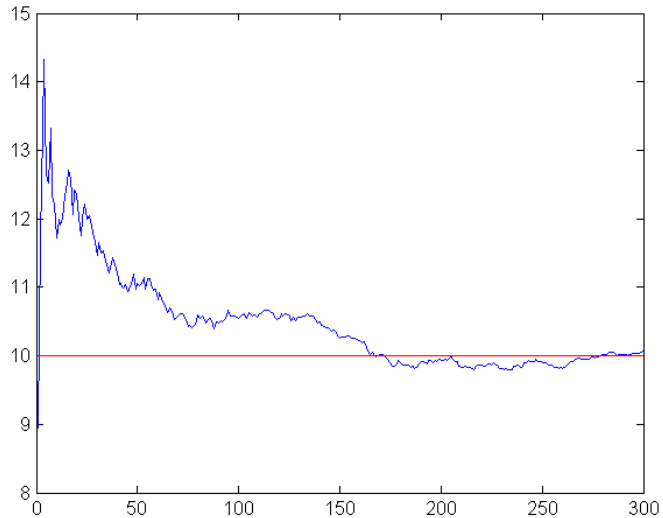
2) approximate m by $\overline{x}_k = 1/k \sum_{i=1}^{k} x_i$



- unbiased estimate since $E(\overline{x}_k) = m$,
- consistant estimate since $Var(\overline{x}_k) \longrightarrow 0$.
- Thus, $\overline{x}_k \longrightarrow m$ when $k \longrightarrow \infty$.

$$\overline{x}_k = \overline{x}_{k-1} + \frac{1}{k}(x_k - \overline{x}_{k-1})$$

2) approximate m by $\overline{x}_k = 1/k \sum_{i=1}^{k} x_i$



- unbiased estimate since $E(\overline{x}_k) = m$,
- consistant estimate since $Var(\overline{x}_k) \rightarrow 0$.
- Thus, $\overline{x}_k \rightarrow m$ when $k \rightarrow \infty$.

$$\overline{x}_k = \overline{x}_{k-1} + \frac{1}{k}(x_k - \overline{x}_{k-1})$$

The SAEM algorithm (Stochastic Approximation of EM)

Delyon, Lavielle and Moulines (the Annals of Statistics, 1999)

First stage of the algorithm:

Iteration k of the algorithm:

- step E :
  - *Simulation:* draw the non observed data $\phi^{(k)}$ with the conditional distribution $p(\phi \,|y; \theta_{k-1})$

- step M: update the estimation of θ:

$$\theta_k = Argmax \;\; p(y, \phi^{(k)}; \theta)$$

The SAEM algorithm (Stochastic Approximation of EM)

Delyon, Lavielle and Moulines (the Annals of Statistics, 1999)

First stage of the algorithm:

Iteration k of the algorithm:

- step E :
  - *Simulation:* draw the non observed data $\phi^{(k)}$ with the conditional distribution $p(\phi \,|y;\theta_{k-1})$
- step M: update the estimation of θ:

$$\theta_k = Argmax \;\; p(y, \phi^{(k)}; \theta)$$

The SAEM algorithm (Stochastic Approximation of EM)

Delyon, Lavielle and Moulines (the Annals of Statistics, 1999)

Second stage of the algorithm:

Iteration k of the algorithm:

- step E :
  - *Simulation:* draw the non observed data $\phi^{(k)}$ with the conditional distribution $p(\phi \,|y;\theta_{k-1})$
  - *Stochastic approximation:*

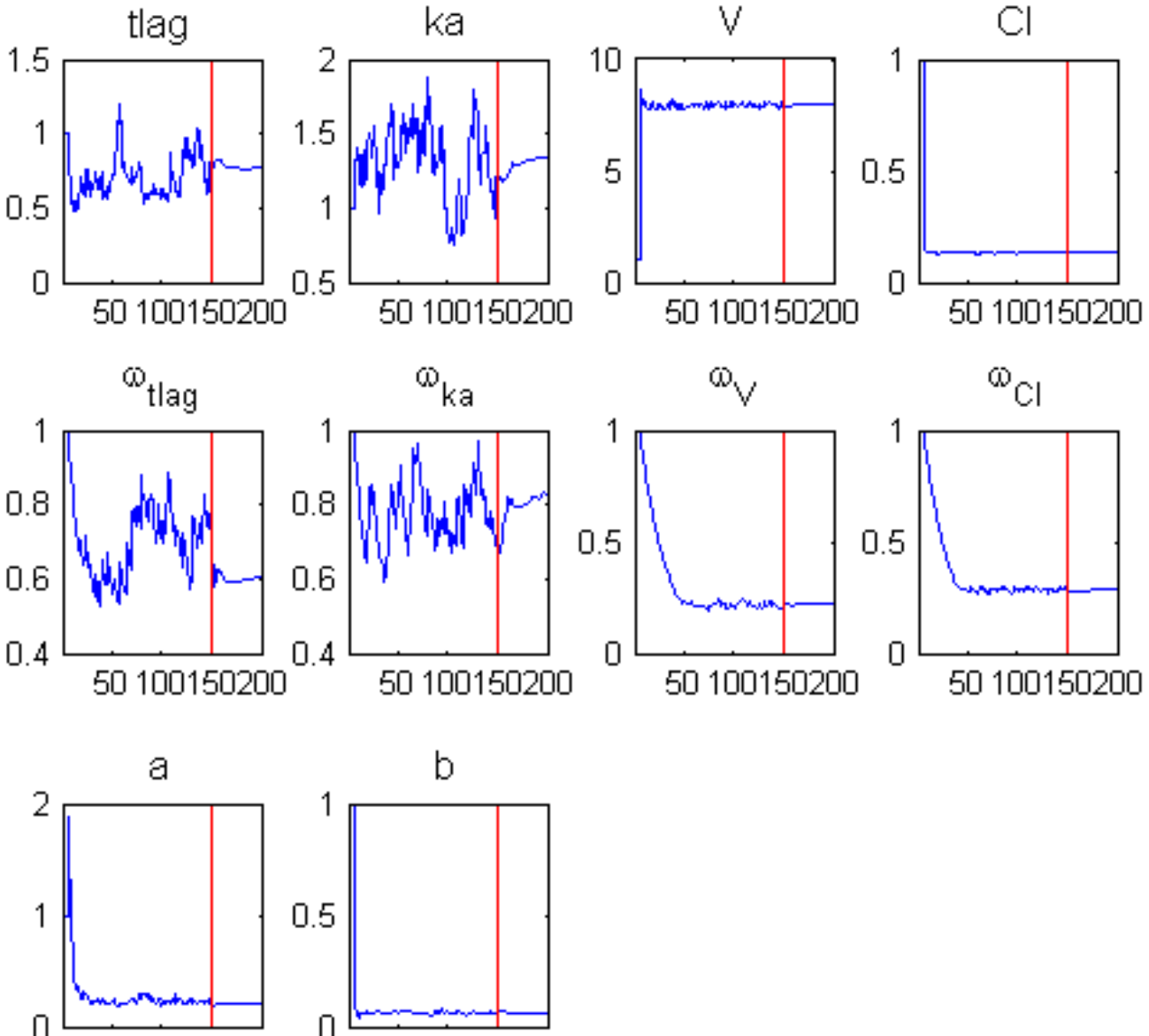
$$\boxed{Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\log p(y, \phi^{(k)};\theta) - Q_{k-1}(\theta)\right]}$$

  (γ_k) is a decreasing sequence: $\sum \gamma_k = +\infty$, $\sum \gamma_k^2 < +\infty$.

- step M: update the estimation of θ:

$$\boxed{\theta_k = Argmax \;\; Q_k(\theta)}$$

The SAEM algorithm (Stochastic Approximation of EM)



Coupling SAEM with MCMC

Kuhn and Lavielle, ESAIM P&S, 2004

Let Π_θ be the transition probability of an *ergodic Markov Chain* with limiting distribution $p_{\Phi|Y}(\cdot|y;\theta)$.

Iteration k of the algorithm:

- *Simulation* : draw $\phi^{(k)}$ according to the transition probability $\Pi_{\theta_{k-1}}(\phi^{(k-1)},\cdot)$.
- *Stochastic approximation:*

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\log p(y,\phi^{(k)};\theta) - Q_{k-1}(\theta)\right]$$

- *Maximization:*

$$\theta_k = Argmax \;\; Q_k(\theta)$$

Coupling SAEM with MCMC

Kuhn and Lavielle, ESAIM P&S, 2004

Let Π_θ be the transition probability of an *ergodic Markov Chain* with limiting distribution $p_{\Phi|Y}(\cdot|y;\theta)$.

Iteration k of the algorithm:

- *Simulation* : draw $\phi^{(k)}$ according to the transition probability $\Pi_{\theta_{k-1}}(\phi^{(k-1)},\cdot)$.
- *Stochastic approximation:*

$$\boxed{Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\log p(y,\phi^{(k)};\theta) - Q_{k-1}(\theta)\right]}$$

- *Maximization:*

$$\boxed{\theta_k = Argmax \;\; Q_k(\theta)}$$

The main convergence Theorem

Theorem

Under very general technical conditions, the SAEM sequence (θ_k) *converges a.s. to some (local) maximum of the observed likelihood* $g(y;\theta)$.

Proof.

See

1. Delyon, Lavielle & Moulines *The Annals of Statistics* (1999)
2. Kuhn & Lavielle *ESAIM P&S* (2004)

Convergence of the algorithm
(Kuhn & Lavielle, 2004)

C1 The chain $(\phi_k)_{k\geq 0}$ takes its values in a compact subset $\mathcal{E}$ of $\mathbb{R}^l$.

C2 For any compact subset V of Θ, there exists a real constant L such that for any (θ, θ') in V^2

$$\sup_{(x,y)\in\mathcal{E}^2} |\Pi_\theta(x,y) - \Pi_{\theta'}(x,y)| \leq L|\theta - \theta'|.$$

C3 The transition probability Π_θ generates a uniformly ergodic chain whose invariant probability is $p(\cdot|y;\theta)$: there exists $K_\theta \in \mathbb{R}+$ and $\rho_\theta \in]0,1[$ such that

$$\forall \phi \in \mathcal{E},\ \forall k \in \mathbb{N}, \qquad ||\Pi_\theta^k(\phi,\cdot) - p(\cdot|y;\theta)||_{TV} \leq K_\theta \rho_\theta^k,$$

$$K \triangleq \sup_\theta K_\theta < +\infty \quad \text{and} \quad \rho \triangleq \sup_\theta \rho_\theta < 1.$$

Convergence of the algorithm
(Kuhn & Lavielle, 2004)

Theorem

- Assume that the regularity conditions required for the convergence of EM are satisfied
- Assume that assumptions C1-C3 hold
- Assume that for any $\theta \in \Theta$, the sequence $(Q_k(\theta))_{k \geq 0}$ takes its values in a compact subset of $\mathcal{S}$.

Then, w.p. 1, $\lim_{k \to +\infty} d(\theta_k, \mathcal{L}) = 0$ where $d(x, A)$ denotes the distance of x to the closed subset A and
$\mathcal{L} = \{\theta \in \Theta, \partial_\theta g(y; \theta) = 0\}$ is the set of stationary points of g.

(Some weak hypothesis ensure the convergence to a (local) maximum of the likelihood)

Estimation of the Fisher Information matrix

An estimate of the asymptotic covariance matrix of $\widehat{\theta}_\ell$ is the inverse of the observed Fisher Information matrix :

$$-\partial_\theta^2 \log g(y; \widehat{\theta}_\ell)$$

Louis's missing information principle (1982)

$$\partial_\theta^2 \log g(y;\theta) = E_{y;\theta}[\partial_\theta^2 \log f(y, Z;\theta)] + \mathrm{Cov}_{y;\theta}[\partial_\theta \log f(y, Z;\theta)]$$

where

$$\begin{aligned}\mathrm{Cov}_{y;\theta}[\partial_\theta \log f(y, Z;\theta)] = {} & E_{y;\theta}[\left(\partial_\theta \log f(y, Z;\theta)\right)\left(\partial_\theta \log f(y, Z;\theta)\right)'] \\ & - E_{y;\theta}[\partial_\theta \log f(y, Z;\theta)] E_{y;\theta}[\partial_\theta \log f(y, Z;\theta)]'\end{aligned}$$

and

$$\partial_\theta \log g(y;\theta) = E_{y;\theta}[\partial_\theta \log f(y, Z;\theta)]$$

Estimation of the Fisher Information matrix

Stochastic approximation:

$$\Delta_k = \Delta_{k-1} + \gamma_k[\partial_\theta \log f(y, \phi_k; \theta_k) - \Delta_{k-1}]$$
$$D_k = D_{k-1} + \gamma_k \left[\partial^2_\theta \log f(y, \phi_k; \theta_k) - Dk - 1\right]$$
$$G_k = G_{k-1} + \gamma_k \left[\partial_\theta \log f(y, \phi_k; \theta_k) \partial_\theta \log f(y, \phi_k; \theta_k)^t - G_{k-1}\right]$$
$$H_k = D_k + G_k - \Delta_k \Delta_k^t$$

Under some regularity conditions, the sequence (H_k) converges almost surely to the Fisher Information matrix

MCMC (Markov Chain Monte Carlo)

An iterative procedure for the simulation of $p(\phi|y;\theta)$

At iteration ℓ

1. draw a new value ϕ^c with a *proposal distribution* q,
2. accept this new value, that is set $\phi^\ell = \phi^c$ with probability

$$\alpha(\phi^c) = \frac{q(\phi^c, \phi^{(\ell-1)})p(\phi^c|y;\theta)}{q(\phi^c, \phi^{(\ell-1)})p(\phi^{(\ell-1)}|y;\theta)}$$

In the model

$$y = f(x;\phi) + g(x;\phi)\varepsilon,$$

computing $\alpha(\phi^c)$ only requires to compute $f(x,\phi^c)$ and $g(x,\phi^c)$ but not the derivatives of f and g

Three following proposal kernels for $1 \leq i \leq N$:

1. $q_{\theta_k}^{(1)}$ is the prior distribution of ϕ_i at iteration k, that is the Gaussian distribution $\mathcal{N}(A_i \mu_k, \Gamma_k)$
2. $q_{\theta_k}^{(2)}$ is the multidimensional random walk $\mathcal{N}(\phi_i, \tau_k \Gamma_k)$.

$$\tau_k = \tau_{k-1}(1 + a(\overline{\rho}_{k-1} - \rho^\star)) \quad 0 < a < 1; \rho^\star \approx 0.4$$

   $\overline{\rho}_{k-1}$: proportion of acceptation at iteration $k-1$.

3. $q_{\theta_k}^{(3)}$ is a succession of d unidimensional Gaussian random walks: each component of ϕ_i are successively updated.

Then, the simulation-step at iteration k consists in running

1. m_1 iterations of the Hasting-Metropolis with proposal $q_{\theta_k}^{(1)}$,
2. m_2 iterations with proposal $q_{\theta_k}^{(2)}$
3. m_3 iterations with proposal $q_{\theta_k}^{(3)}$.

A Simulated Annealing version of SAEM

Conditional distribution of ϕ:

$$p_{\Phi|Y}(\ \phi\ |y;\theta) = C(y;\theta)e^{-U(\phi,y;\theta)}$$

Temperature parameter T:

$$p^{(T)}_{\Phi|Y}(\ \phi\ |y;\theta) = C_T(y;\theta)e^{-\frac{U(\phi,y;\theta)}{T}}$$

Choose a decreasing Temperature sequence (T_k) converging to 1. Then, at iteration k of SAEM,

- E-step: draw the non observed data $\phi^{(k)}$ with the conditional distribution $p^{(T_k)}_{\Phi|Y}(\ \cdot\ |y;\theta_{k-1})$
- M-step remains unchanged

Importance Sampling for estimating the marginal likelihood

The Importance Sampling algorithm computes an estimate $\ell_M(y)$ of the observed likelihood.

$$\begin{aligned}
\ell(y,\theta) &= \int p(y,\phi)d\phi \\
&= \int h(y|\phi)\pi(\phi)d\phi \\
&= \int \left(h(y|\phi)\frac{\pi(\phi)}{\tilde{\pi}(\phi)} \right) \tilde{\pi}(\phi)d\phi
\end{aligned}$$

1. Draw $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(M)}$ with the distribution $\tilde{\pi}$,
2. Let

$$\ell_M(y) = \frac{1}{M}\sum_{j=1}^{M} h(y|\phi^{(j)})\frac{\pi(\phi^{(j)})}{\tilde{\pi}(\phi^{(j)})}$$

Importance Sampling for estimating the marginal likelihood

$$\ell_M(y) = \frac{1}{M}\sum_{j=1}^{M} h(y|\phi^{(j)})\frac{\pi(\phi^{(j)})}{\tilde{\pi}(\phi^{(j)})}$$

$\mathrm{E}\ell_M(y) = \ell(y)$ and $\mathrm{Var}\ell_M(y) = \mathcal{O}(1/M)$

The instrumental distribution used in MONOLIX:

1) Estimate the conditional mean and variance of ϕ,
2) Use for $\tilde{\pi}$ a decentred t-distribution with ν d.f.
   $\phi_i^{(j)} = E(\phi_i|y_i;\widehat{\theta}) + s.d.(\phi_i|y_i;\widehat{\theta}) \times T(\nu)$