

Supporting slides

Mohamed Tarek

¹PumasAI Inc.

²University of Sydney Business School

Section 1

NLME fitting algorithms

Notation

- θ : all the population parameters (typical values, BSV, BOV and residual variability)

Notation

- θ : all the population parameters (typical values, BSV, BOV and residual variability)
- η_i : random effects of subject i

Notation

- θ : all the population parameters (typical values, BSV, BOV and residual variability)
- η_i : random effects of subject i
- N : number of subjects

Notation

- θ : all the population parameters (typical values, BSV, BOV and residual variability)
- η_i : random effects of subject i
- N : number of subjects
- $\eta = \{\eta_i, \forall i \in 1 \dots N\}$: random effects of all the subjects

Notation

- θ : all the population parameters (typical values, BSV, BOV and residual variability)
- η_i : random effects of subject i
- N : number of subjects
- $\eta = \{\eta_i, \forall i \in 1 \dots N\}$: random effects of all the subjects
- y_i : observations of subject i

Notation

- θ : all the population parameters (typical values, BSV, BOV and residual variability)
- η_i : random effects of subject i
- N : number of subjects
- $\eta = \{\eta_i, \forall i \in 1 \dots N\}$: random effects of all the subjects
- y_i : observations of subject i
- x_i : covariates of subject i

NLME Notation

- $p(y = y_i \mid \theta = \theta, \eta = \eta_i, x = x_i) = p(y_i \mid \theta, \eta_i, x_i)$: likelihood of (θ, η_i) given subject i 's data (x_i, y_i) . Also known as the conditional probability of y_i given θ, η_i, x_i . Or just the conditional likelihood of η_i given θ and (x_i, y_i) .

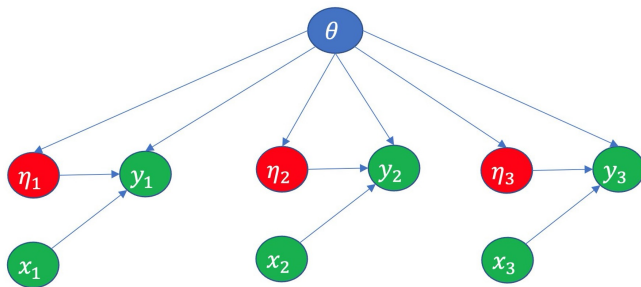
NLME Notation

- $p(y = y_i \mid \theta = \theta, \eta = \eta_i, x = x_i) = p(y_i \mid \theta, \eta_i, x_i)$: likelihood of (θ, η_i) given subject i 's data (x_i, y_i) . Also known as the conditional probability of y_i given θ, η_i, x_i . Or just the conditional likelihood of η_i given θ and (x_i, y_i) .
- $p(\eta = \eta_i \mid \theta = \theta) = p(\eta_i \mid \theta)$: prior probability of the random effects η_i given the population parameters θ .

NLME Notation

- $p(y = y_i \mid \theta = \theta, \eta = \eta_i, x = x_i) = p(y_i \mid \theta, \eta_i, x_i)$: likelihood of (θ, η_i) given subject i 's data (x_i, y_i) . Also known as the conditional probability of y_i given θ, η_i, x_i . Or just the conditional likelihood of η_i given θ and (x_i, y_i) .
- $p(\eta = \eta_i \mid \theta = \theta) = p(\eta_i \mid \theta)$: prior probability of the random effects η_i given the population parameters θ .
- $p(y = y_i \mid \theta = \theta, x = x_i) = p(y_i \mid \theta, x_i) = \int p(y_i \mid \theta, \eta_i, x_i) \cdot p(\eta_i \mid \theta) d\eta_i$: marginal likelihood of θ given subject i 's data y_i .

NLME Notation



NLME Notation

```

@model begin
  @param begin
     $\theta$   $\in$  VectorDomain(4, lower = zeros(4))
     $\Omega$   $\in$  PSDDomain(2)
     $\Sigma$   $\in$  RealDomain(lower = 0.0)
     $a$   $\in$  RealDomain(lower = 0.0, upper = 1.0)
  end
  @random begin
     $\eta_i | \theta$   $\eta \sim \text{MvNormal}(\Omega)$ 
  end
  @covariates sex wt etn
   $x_i$ 
  @pre begin
     $\theta_1 := \theta[1]$ 
    Ka =  $\theta_1$ 
    CL =  $\theta[2] * ((wt / 70)^{0.75} * (\theta[4]^{\text{sex}} * \exp(\eta[1])))$ 
    Vc =  $\theta[3] * \exp(\eta[2])$ 
  end
   $y_i | \theta, \eta_i, x_i$ 
  @dynamics begin
    Depot' = -Ka * Depot
    Central' = Ka * Depot - (CL / Vc) * Central
    Res' = Depot - Central
  end
  @derived begin
    conc = @. Central / Vc
    dv ~ @. Normal(conc, conc *  $\Sigma$ )
    T_max = maximum(t)
  end
  @observed begin
    obs_cmax = maximum(dv)
  end
end

```

Marginal likelihood maximization

$$\begin{aligned}\theta^* &= \arg \max_{\theta} \prod_{i=1}^N p(y_i \mid \theta, x_i) \\ &= \arg \max_{\theta} \prod_{i=1}^N \int p(y_i \mid \theta, \eta_i, x_i) \cdot p(\eta_i \mid \theta) d\eta_i \\ \text{EBE}_i &= \eta_i^* = \arg \max_{\eta_i} \left(p(y_i \mid \theta = \theta^*, \eta_i, x_i) \cdot p(\eta_i \mid \theta = \theta^*) \right)\end{aligned}$$

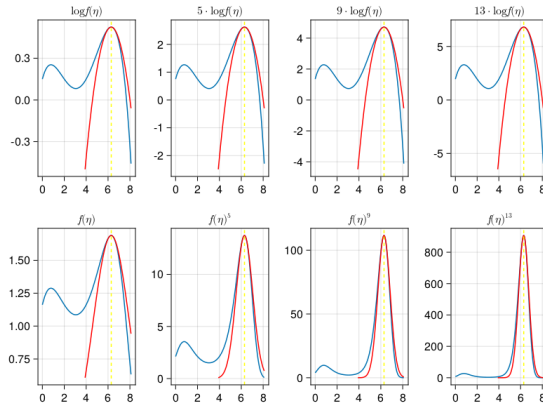
General Laplace method

$$\int f(\eta) d\eta \approx f(\eta^*) \sqrt{(2\pi)^m / | - H |}$$

- f : a positive scalar-valued function of η
- η : vector of m integration variables
- η^* : global maximizer of $\log f$, $\frac{d \log f}{d\eta}(\eta^*) = 0$
- H : second derivative matrix of $\log f$ wrt η at η^* , must be negative definite at $\eta = \eta^*$
- $| - H |$: determinant of $-H$

NLME Laplace method

Laplace uses a second order Taylor series approximation of $\log f$ at η^* .



NLME Laplace method

The mode is the same after multiplying $\log f$ by a scalar. The ratio of peaks is not the same. Consider the 2 local maximizers η_1 (lower peak) and η_2 (higher peak).

$$\begin{aligned}c &= \log f(\eta_2) - \log f(\eta_1) \\n \cdot c &= n \cdot (\log f(\eta_2) - \log f(\eta_1)) \\e^{n \cdot c} &= f(\eta_2)^n / f(\eta_1)^n\end{aligned}$$

Summary

Approximation error of $n \cdot \log f$ away from the mode η^* is not significant as n increases.

NLME Laplace method

- There are N functions $\{f_i : i = 1 \dots N\}$ to be integrated, one for each subject
- $f_i(\eta_i) = p(y_i \mid \theta, \eta_i, x_i) \cdot p(\eta_i \mid \theta)$
- Laplace method

$$\int f_i(\eta_i) d\eta_i = f_i(\text{EBE}_i) \sqrt{(2\pi)^m / | - H_i |}$$

- H_i : second derivative matrix of $\log f_i$ wrt η_i at EBE_i , must be negative definite at $\eta_i = \text{EBE}_i$

General FOCE(I)

FOCE(I) approximates the Hessian H_i for each subject i . Assume the following:

$$\begin{aligned}\log f_i(\eta_i) &= \log p(y_i \mid \theta, \eta_i, x_i) + \log p(\eta_i \mid \theta) \\ &= L_i(g_i(\eta_i)) + \log p(\eta_i \mid \theta)\end{aligned}$$

where:

- g_i returns the vector of IPREDs μ_i and the residual standard deviations σ_i (constant in the additive error model case), at all observed time points, and
- L_i is the log probability of y_i given μ_i and σ_i

General FOCE(I)

g_i is usually the most expensive component of $\log f_i$, because it often involves solving a differential equation. So let's approximate it!

First order Taylor series expansion

- FO

$$g_i(\eta_i) = g_i(0) + \frac{dg_i}{d\eta_i}(0) \cdot \eta_i$$

- FOCE(I)

$$g_i(\eta_i) = g_i(\text{EBE}_i) + \frac{dg_i}{d\eta_i}(\text{EBE}_i) \cdot (\eta_i - \text{EBE}_i)$$

General FOCE(I)

Summary

- FOCE(I) ensures that the approximation error in g_i (and $\log f_i$ by extension) is low in the proximity of EBE_i .
- FO does not ensure that so it only works well if:
 - EBE_i is not far from 0, or
 - g_i is close to linear in the interval $[0, \text{EBE}_i]$.
- FO requires a correction term in the Laplace method because the gradient of $\log f_i$ wrt η_i at $\eta_i = 0$ is not 0.

General FOCE(I)

Chain rule for Hessians

$$(L_i \cdot g_i)''(\eta_i) = \frac{dg_i^T}{d\eta_i} \cdot \frac{\partial^2 L_i}{\partial g_i \cdot \partial g_i^T} \cdot \frac{dg_i}{d\eta_i} + \sum_{t=1}^d \left(\frac{\partial L_i}{\partial g_{i,t}} \cdot \underbrace{\frac{\partial^2 g_{i,t}}{\partial \eta_i \cdot \partial \eta_i^T}}_{0 \text{ if linear}} \right)$$

where d is twice the number of observed time points (corresponding to μ_i and σ_i) and $g_{i,t}$ is the t^{th} component of g_i .

General FOCE(I)

Summary

If g_i is linear in η_i :

- $\frac{\partial^2 g_{i,t}}{\partial \eta_i \cdot \partial \eta_i^T} = 0$
- $J_i = \frac{\partial g_i}{\partial \eta_i}$ is constant

The Hessian simplifies to:

$$(L_i \cdot g_i)''(\eta_i) = J_i^T \cdot \frac{\partial^2 L_i}{\partial g_i \cdot \partial g_i^T} \cdot J_i$$

General FOCE(I)

- J_i can be computed for each subject using finite difference at
 - $\eta_i = 0$ for FO, or
 - $\eta_i = \text{EBE}_i$ for FOCE(I)
- For many data distributions, $\frac{\partial^2 L_i}{\partial g_i \cdot \partial g_i^T}$ is both diagonal and has a closed form. Doesn't have to be Gaussian!

General FOCE(I)

Recall

$$\begin{aligned}\log f_i(\eta_i) &= \log p(y_i \mid \theta, \eta_i, x_i) + \log p(\eta_i \mid \theta) \\ &= L_i(\underbrace{g_i(\eta_i)}_{\text{approx}}) + \log p(\eta_i \mid \theta)\end{aligned}$$

For many random effects distributions, the Hessian of $\log p(\eta_i \mid \theta)$ wrt η_i has a closed form. Doesn't have to be Gaussian!

General FOCE(I)

Summary

- In Pumas, FOCE is always “with interaction”.
- Pumas FOCE supports a number of data distributions:
 - Continuous: Normal, LogNormal, Gamma, Exponential, Beta
 - Discrete: NegativeBinomial, Bernoulli, Binomial, Poisson, Categorical

General FOCE(I)

Summary

- Pumas supports a number of random effect distributions:
 - **Unbounded:** Cauchy, Gumbel, Laplace, Logistic, Normal, NormalCanon, NormalInverseGaussian, PGeneralizedGaussian, TDist
 - **Positive:** BetaPrime, Chi, Chisq, Erlang, Exponential, Frechet, Gamma, InverseGamma, InverseGaussian, Kolmogorov, LogNormal, NoncentralChisq, Rayleigh, Weibull
 - **Between 0 and 1:** Beta, LogitNormal
 - **Other bounded:** Uniform, Arcsine, Biweight, Cosine, Epanechnikov, LogUniform, Semicircle, SymTriangularDist, Triweight

General FOCE(I)

Summary

- Use FOCE if supported, otherwise use Laplace.
- Avoid FO.

Section 2

Diagnostics

Standard error estimation

Goal

Estimate the covariance matrix of the estimator θ^* :

$$\theta^* = \arg \max_{\theta} \prod_{i=1}^N p(y_i \mid \theta, x_i)$$

Sandwich estimator of standard errors

Asymptotic covariance

$$\theta^* \sim \mathcal{N}(\theta_0, V) = \mathcal{N}(\theta_0, A^{-1}BA^{-1})$$

$$A = \sum_{i=1}^N \frac{\partial^2 \log p(y_i \mid \theta, x_i)}{\partial \theta \cdot \partial \theta^T}(\theta_0)$$

$$B = \sum_{i=1}^N \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta}(\theta_0) \times \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta}(\theta_0)^T$$

where θ_0 is the set of unknown true parameters.

Sandwich estimator of standard errors

Estimated covariance

$$\theta^* \stackrel{a}{\sim} \mathcal{N}(\theta_0, \hat{V}) = \mathcal{N}(\theta_0, \hat{A}^{-1} \hat{B} \hat{A}^{-1})$$

$$\hat{A} = \sum_{i=1}^N \frac{\partial^2 \log p(y_i \mid \theta, x_i)}{\partial \theta \cdot \partial \theta^T}(\theta^*)$$

$$\hat{B} = \sum_{i=1}^N \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta}(\theta^*) \times \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta}(\theta^*)^T$$

Sandwich estimator of standard errors

Standard error estimates

The square root of the diagonal elements of \hat{V} are the standard error estimates of θ^* .

Sandwich estimator of standard errors

Computing \hat{V}

- \hat{A} and \hat{B} can be approximated with finite difference
- $\log p(y_i | \theta, x_i)$ itself needs to be approximated with Laplace/FOCE/FO
- $\hat{A}^{-1} \cdot \hat{B} \cdot \hat{A}^{-1}$ is computed using a generalized eigenvalue problem.

Sandwich estimator of standard errors

Computing \hat{V}

Consider the following generalized eigenvalue problem:

$$\begin{aligned}\hat{B} \cdot U &= \hat{A} \cdot U \cdot \Lambda \\ I &= U^T \cdot \hat{A} \cdot U\end{aligned}$$

where U is the matrix of generalized eigenvectors and Λ is the diagonal matrix of generalized eigenvalues.

Sandwich estimator of standard errors

Computing \hat{V}

The inverse of the matrix of eigenvectors U is obtained from the following constraint on U :

$$(U^T \cdot \hat{A}) \cdot U = I$$
$$U^{-1} = U^T \cdot \hat{A}$$

Sandwich estimator of standard errors

Computing \hat{V}

The following identity is true:

$$\hat{B} \cdot U = \hat{A} \cdot U \cdot \Lambda$$

$$\hat{A}^{-1} \cdot \hat{B} = U \cdot \Lambda \cdot U^{-1}$$

$$\hat{A}^{-1} \cdot \hat{B} = U \cdot \Lambda \cdot U^T \cdot \hat{A}$$

$$\hat{V} = \hat{A}^{-1} \cdot \hat{B} \cdot \hat{A}^{-1} = U \cdot \Lambda \cdot U^T$$

Sandwich estimator of standard errors

Failed estimator

- If the computed \hat{A} is: a) singular, b) near singular, or c) has negative eigenvalues, the sandwich estimator will fail.
- This is a sign of poor identifiability of at least 1 parameter and/or significant numerical errors.
- Even if a single (IIV) parameter is not identifiable given the data, \hat{A} will be singular.

Sandwich estimator of standard errors

Failed estimator

- Numerical errors in the finite difference or Laplace/FOCE/FO can also cause the computed approximate \hat{A} to be singular (or have small negative eigenvalues) even when the exact matrix \hat{A} may be only *near* singular and positive definite.

Weighted residuals

Visual predictive check

Section 3

Time to event

Definitions

Quantity	Formula
Instantaneous hazard	$\lambda(t) > 0$
Cumulative hazard	$\Lambda(t) = \int_0^t \lambda(t') dt'$
Survival function: probability of survival up to time t	$S(t) = \exp(-\Lambda(t))$
Failure function: probability of death/failure before time t	$F(t) = 1 - S(t)$
Probability density function of time of death t	$f(t) = \frac{dF}{dt} = \lambda(t) \cdot \exp(-\Lambda(t))$
Expected time of death $E[t]$	$E[t] = \int_0^\infty t \cdot f(t) dt = \int_0^\infty S(t) dt$

Log likelihood

The log likelihood for censored survival data is given by the following 2 formulas:

- For censored subjects at time t (patient survived until time t)

$$\log \text{likelihood} = \log S(t) = -\Lambda(t)$$

- For subjects dead at time t

$$\log \text{likelihood} = \log f(t) = \log \lambda(t) - \Lambda(t)$$