# PAGANZ 2024 Pumas Workshop

Mohamed Tarek

[1]PumasAI Inc.

[2]University of Sydney Business School

# NLME fitting algorithms

# Notation

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

# Notation

- $\theta$: all the population parameters (typical values, BSV, BOV and residual variability)
- $\eta_i$: random effects of subject $i$

# Notation

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

# Notation

- $\theta$: all the population parameters (typical values, BSV, BOV and residual variability)
- $\eta_i$: 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

- $\theta$: all the population parameters (typical values, BSV, BOV and residual variability)
- $\eta_i$: 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

- $\theta$: all the population parameters (typical values, BSV, BOV and residual variability)
- $\eta_i$: 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 $(\theta, \eta_i)$ given subject $i$'s data $(x_i, y_i)$. Also known as the conditional probability of $y_i$ given $\theta, \eta_i, x_i$. Or just the conditional likelihood of $\eta_i$ given $\theta$ and $(x_i, y_i)$.

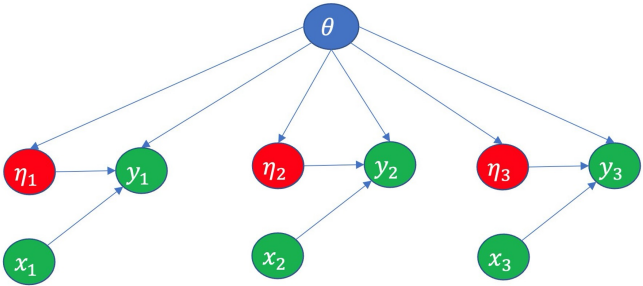
# NLME Notation

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

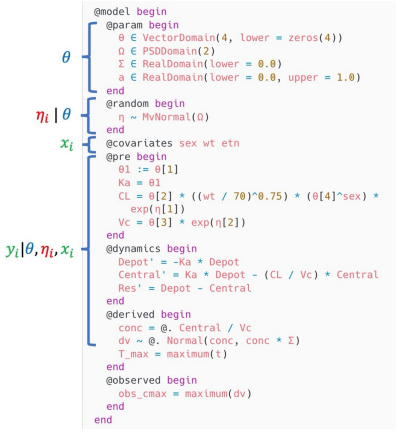
# NLME Notation

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

# NLME Notation



# NLME Notation




PumasAI

# Marginal likelihood maximization

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

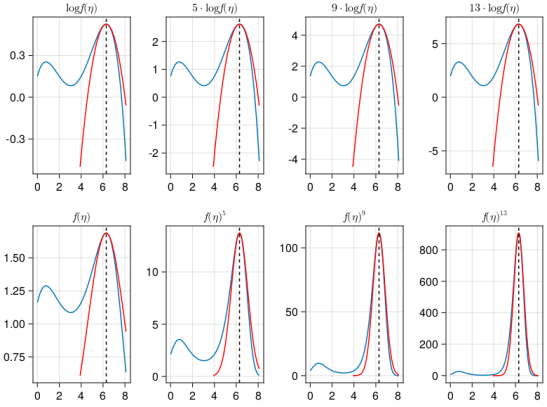
# General Laplace method

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

- $f$: a positive scalar-valued functon of $\eta$
- $\eta$: vector of $m$ integration variables
- $\eta^*$: global maximizer of $\log f$, $\frac{d \log f}{d\eta}(\eta^*) = 0$
- $H$: second derivative matrix of $\log f$ wrt $\eta$ at $\eta^*$, 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 $\eta^*$.




PumasAI

# NLME Laplace method

Consider the 2 local maximizers $\eta_1$ (lower peak) and $\eta_2$ (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 $\eta^*$ 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 $\eta_i$ at $\text{EBE}_i$, 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 $\mu_i$ and the residual standard deviations $\sigma_i$ (constant in the additive error model case), at all observed time points, and
- $L_i$ is the log probability of $y_i$ given $\mu_i$ and $\sigma_i$

# 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)

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

# 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 $\text{EBE}_i$.
- FO does not ensure that so it only works well if:
  - $\text{EBE}_i$ 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 $\eta_i$ at $\eta_i = 0$ is not 0.

# General FOCE(I)

## Chain rule for Hessians

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

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

# General FOCE(I)

## Summary

If $g_i$ is linear in $\eta_i$:

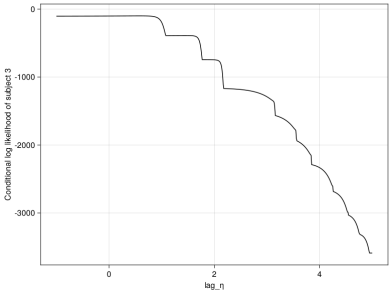
- $\frac{\partial^2 g_{i,t}}{\partial \eta_i \cdot \partial \eta_i^T} = 0$
- $J_i = \frac{\partial g_i}{\partial \eta_i}$ is constant

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)

- One surprising advantage of FOCE(I) is that the Hessian approximation is often negative definite even when the exact Hessian is singular or not well defined at $\eta_i = \eta_i^*$.



# 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 $\eta_i$ has a closed form. Doesn't have to be Gaussian!

# General FOCE(I)

- Pumas FOCE supports a number of data distributions:
  - **Continuous**: Normal, LogNormal, Gamma, Exponential, Beta
  - **Discrete**: NegativeBinomial, Bernoulli, Binomial, Poisson, Categorical

# General FOCE(I)

- 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

- In Pumas, FOCE is always "with interaction".
- Use FOCE if supported, otherwise use Laplace.
- Avoid FO.

# Standard error estimation

# Goal

Estimate the covariance matrix of the estimator $\theta^*$:

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

# 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 $\theta_0$ is the set of unknown true parameters.

# Sandwich estimator of standard errors

## Estimated covariance

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

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

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

# Sandwich estimator of standard errors

## Standard error estiamtes

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

# Sandwich estimator of standard errors

## Computing $\hat{V}$

- $\hat{A}$ and $\hat{B}$ can be approximated with finite difference
- $\log p(y_i \mid \theta, x_i)$ 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:

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

where $U$ is the matrix of generalized eigenvectors and $\Lambda$ 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:

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

# 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 $\widehat{A}$ to be singular (or have small negative eigenvalues) even when the exact matrix $\widehat{A}$ may be only *near* singular and positive definite.

# Weighted residuals

# Visual predictive check

# Continuous VPC Procedure

# Time to event VPC Procedure

1. Simulate a synthetic population a given number of samples (`samples`, default 499). For each subject:
    1. Evaluate the cumulative hazard function $\Lambda$ at `nT` (default 10) time points between `minT` and `maxT`.
    2. Use a cubic spline to interpolate between the $\Lambda$ values.
    3. Use inverse CDF transform sampling to sample the time of death from the cumulative hazard function.

# Time to event VPC Procedure

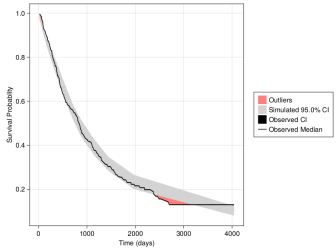
2. Stratify the observed and simulated populations by the stratification variable.
3. For each simulated population stratum:
   1. Estimate the Kaplan Meier (KM) curve. $d_i$ is the number of deaths at $t_i$ and $n_i$ is the number of people at risk at time $t_i$.

$$\hat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right)$$

   2. Combine all simulated populations' KM curves into one data frame.
   3. Do quantile regression with smoothing to get smooth curves for the quantiles at a number of nodes `nnodes` (default 11).

# Time to event VPC Procedure

4. For the observed population stratum, estimate the KM curve.
5. Plot the observed KM curve against the smoothed quantiles.




PumasAI

# Time to event VPC Procedure

## Inverse CDF sampling

- If $R \sim \text{Uniform}(0, 1)$, then $-\log(1 - R) \sim \text{Exponential}(1.0)$.

$$
\begin{aligned}
F(t) &\leq R \\
1 - S(t) &\leq R \\
\exp(-\Lambda(t)) &\geq 1 - R \\
\Lambda(t) &\leq -\log(1 - R)
\end{aligned}
$$

- The sample $t$ is obtained using a root finding algorithm to find the root for $\Lambda(t) = -\log(1 - R)$.

# Time to event models

# Definitions

- Instantaneous hazard

$$\lambda(t) > 0$$

- Cumulative hazard

$$\Lambda(t) = \int_0^t \lambda(t')\, dt'$$

- Survival function: probability of survival up to time $t$

$$S(t) = \exp(-\Lambda(t))$$

# Definitions

- 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 $t$

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