PAGANZ 2024 Pumas Workshop

Mohamed Tarek

¹PumasAl Inc.

²University of Sydney Business School



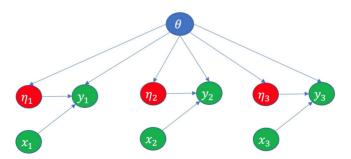




NLME fitting algorithms



Assume there are 3 subjects



```
@model begin
                   @param begin
                     θ € VectorDomain(4, lower = zeros(4))
                     Σ ∈ RealDomain(lower = 0.0)
                     a ∈ RealDomain(lower = 0.0, upper = 1.0)
                   @random begin
      \eta_i \mid \theta
                     n ~ MvNormal(Ω)
                   @covariates sex wt etn
                   @pre begin
                     \theta 1 := \theta [1]
                     CL = \theta[2] * ((wt / 70)^0.75) * (\theta[4]^sex) *
                     V_C = 0.031 * exp(n.021)
                   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 * Σ)
                     T \max = \max \operatorname{imum}(t)
                   end
                   @observed begin
                     obs cmax = maximum(dv)
                   end
                 end
```

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

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

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

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

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

- θ : all the population parameters (typical values, BSV, BOV and residual variability)
- lacksquare 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
- $lackbox{} y_i$: observations of subject i
- \mathbf{x}_i : covariates of subject i

■ $p(y=y_i \mid \theta=\theta, \eta=\eta_i, x=x_i) = p(y_i \mid \theta, \eta_i, x_i)$: 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) .

- $p(y=y_i \mid \theta=\theta, \eta=\eta_i, x=x_i) = p(y_i \mid \theta, \eta_i, x_i)$: 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 .

- $p(y=y_i \mid \theta=\theta, \eta=\eta_i, x=x_i) = p(y_i \mid \theta, \eta_i, x_i)$: 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 .
- $\begin{array}{l} \bullet \ p(y=y_i \mid \theta=\theta, x=x_i) = p(y_i \mid \theta, x_i) = \int p(y_i \mid \theta, \eta_i, x_i) \cdot p(\eta_i \mid \theta) \, d\eta_i \text{: marginal likelihood of } \theta \text{ given subject } i \text{'s data } y_i. \end{array}$



Marginal likelihood maximization

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

General Laplace method

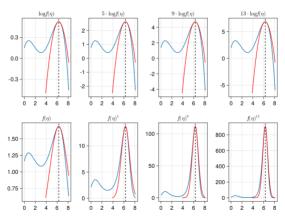
$$\int f(\eta)\,d\eta \approx f(\eta^*)\sqrt{(2\pi)^m/|-H|}$$

- f: a positive scalar-valued function of η
- \blacksquare η : vector of m integration variables
- lacksquare 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

Consider the 2 local maximizers η_1 (lower peak) and η_2 (higher peak).

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

Summary

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



NLME Laplace method

- \blacksquare There are N functions $\{f_i: i=1\dots N\}$ to be integrated, one for each subject
- Laplace method

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

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



 $\mathsf{FOCE}(\mathsf{I})$ approximates the Hessian H_i for each subject i. Assume the following:

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

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
- lacksquare L_i is the log probability of y_i given μ_i and σ_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

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

FOCE(I)

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



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 $[0, \mathsf{EBE}_i]$.
- 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.



Chain rule for Hessians

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

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 .



Summary

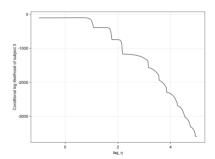
If g_i is linear in η_i :

$$lacksquare J_i = rac{\partial g_i}{\partial \eta_i}$$
 is constant

The Hessian simplifies to:

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

• 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 $\eta_i = \eta_i^*$.



- lacksquare J_i can be computed for each subject using finite difference at
 - \bullet $\eta_i=0$ for FO, or
 - $\eta_i = \mathsf{EBE}_i \text{ for FOCE(I)}$
- 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!

Recall

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

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!

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

- 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



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 θ^* :

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

Asymptotic covariance

$$\begin{split} \theta^* &\sim \mathcal{N}(\theta_0, V) = \mathcal{N}(\theta_0, A^{-1}BA^{-1}) \\ A &= \sum_{i=1}^N \frac{\partial^2 \log p(y_i \mid \theta, x_i)}{\partial \theta \cdot \partial \theta^T}(\theta_0) \\ B &= \sum_{i=1}^N \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta}(\theta_0) \times \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta}(\theta_0)^T \end{split}$$

where θ_0 is the set of unknown true parameters.



Estimated covariance

$$\begin{split} & \theta^* \overset{a}{\sim} \mathcal{N}(\theta_0, \hat{V}) = \mathcal{N}(\theta_0, \hat{A}^{-1} \hat{B} \hat{A}^{-1}) \\ & \hat{A} = \sum_{i=1}^N \frac{\partial^2 \log p(y_i \mid \theta, x_i)}{\partial \theta \cdot \partial \theta^T} (\theta^*) \\ & \hat{B} = \sum_{i=1}^N \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta} (\theta^*) \times \frac{\partial \log p(y_i \mid \theta, x_i)}{\partial \theta} (\theta^*)^T \end{split}$$

Standard error estiamtes

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

Computing \hat{V}

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

Computing \hat{V}

Consider the following generalized eigenvalue problem:

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

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

Computing \hat{V}

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

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

Computing \hat{V}

The following identity is true:

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

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.



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.

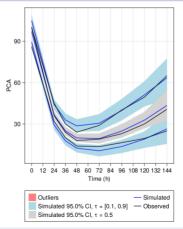
Weighted residuals

Visual predictive check



Continuous VPC Procedure

Example



Continuous VPC Procedure

- Simulate a synthetic population a given number of samples (samples, default 499).
- 2 Stratify the observed and simulated populations by the stratification variable.
- 3 For each simulated population stratum, do smoothed quantile regression at nnodes nodes picked from the the data.
 - Default quantiles: 0.1, 0.5 and 0.9.
 - Default nnodes: 11
 - Default smoothing bandwidth: 2.0



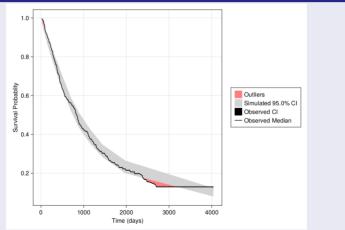


Continuous VPC Procedure

- 4 Find the (hyper-)quantiles of the per-scenario population quantiles within each stratum,
 - Hyper-quantiles:
 - (1 level) / 2
 - 0.5 (simquantile_medians hidden by default)
 - (1 + level) / 2
 - Default level: 0.95
- **5** For each observed population stratum, repeat step 3.
- 6 For each stratum, plot the population's quantiles and the hyper-quantiles of each simulated quantile.



Example



asAl

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



- 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).



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

Inverse CDF sampling

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

$$F(t) \leq R$$

$$1 - S(t) \leq R$$

$$\exp(-\Lambda(t)) \geq 1 - R$$

$$\Lambda(t) \leq -\log(1 - R)$$

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



└─Time to event models

Time to event models



Definitions

Instantaneous hazard

$$\lambda(t) > 0$$

Cumulative hazard

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

lacksquare Survival function: probability of survival up to time t

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

Definitions

lacktriangle Failure function: probability of death/failure before time t

$$F(t) = 1 - S(t)$$

Probability density function of time of death t

$$f(t) = \frac{dF}{dt} = \lambda(t) \cdot \exp(-\Lambda(t))$$

lacksquare 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:

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

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

For subjects dead at time t

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