Sensitivity Equations Provide More Robust Gradients and Faster Computation of the FOCE Approximation to the Population Likelihood

Joachim Almquist[1,2], Jacob Leander[1,3,*], and Mats Jirstrand[1]

[1]Fraunhofer-Chalmers Centre, Göteborg, [2]Department of Chemical and Biological Engineering, Chalmers University of Technology, Göteborg
[3]Department of Mathematical Sciences, Chalmers University of Technology, Göteborg, *Current affiliation AstraZeneca R&D, Mölndal

Background

The first order conditional estimation (FOCE) method [1] is still one of the parameter estimation workhorses for nonlinear mixed effects (NLME) modeling used in population pharmacokinetics and pharmacodynamics [2]. We propose an novel implementation of the FOCE and FOCEI methods where instead of obtaining the gradients needed for the two levels of quasi-Newton optimizations from the standard finite difference approximation, gradients are computed using so called sensitivity equations [3].

The Approximate Population Likelihood

The state-space model for a single individual is described by a system of ordinary differential equations and a corresponding set of measurement equations

$$\frac{d\mathbf{x}_i(t)}{dt} = \mathbf{f}(\mathbf{x}_i(t), t, \mathbf{Z}_i(t), \boldsymbol{\theta}, \boldsymbol{\eta}_i) \qquad \mathbf{y}_{ij} = \mathbf{h}(\mathbf{x}_{ij}, t_{j_i}, \mathbf{Z}_i(t_{j_i}), \boldsymbol{\theta}, \boldsymbol{\eta}_i) + \mathbf{e}_{ij}$$

$$\mathbf{x}_i(t_0) = \mathbf{x}_{0i}(\mathbf{Z}_i(t_0), \boldsymbol{\theta}, \boldsymbol{\eta}_i) \qquad \mathbf{e}_{ij} \in N(\mathbf{0}, \mathbf{R}_{ij}(\mathbf{x}_{ij}, t_{j_i}, \mathbf{Z}_i(t_{j_i}), \boldsymbol{\theta}, \boldsymbol{\eta}_i))$$

$$\hat{\mathbf{y}}_{ij} = \mathrm{E}[\mathbf{y}_{ij}]$$

where indices i and j denote individuals and observations, respectively. Furthermore, θ are fixed effects parameters, $Z_i(t_j)$ are covariates, $\eta_i \sim N(0, \Omega)$ are random effect parameters, and R_{ij} are measurement error covariance matrices.

Given a set of experimental observations, d_{ij}, for the individuals $i = 1, \ldots, N$ at the time points t_{ij}, where $j = 1, \ldots n_i$, we define the residuals $\epsilon_{ij} = \mathrm{d}_{ij} - \hat{\mathrm{y}}_{ij}$

The approximate log-likelihood function is obtained using the Laplacian approximation, which involves a second order Taylor expansion wrt η_i of l_i around points η_i^* that maximizes the individual l_i.

$$\log L(\boldsymbol{\theta}) \approx \log L_F(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(l_i(\boldsymbol{\eta}_i^*) - \frac{1}{2} \log\det\left[\frac{-\mathbf{H}_i(\boldsymbol{\eta}_i^*)}{2\pi}\right]\right)$$

where

$$l_i = -\frac{1}{2}\sum_{j=1}^{n_i}\left(\boldsymbol{\epsilon}_{ij}^T \mathbf{R}_{ij}^{-1}\boldsymbol{\epsilon}_{ij} + \log\det(2\pi \mathbf{R}_{ij})\right) - \frac{1}{2}\boldsymbol{\eta}_i^T \boldsymbol{\Omega}^{-1}\boldsymbol{\eta}_i - \frac{1}{2}\log\det(2\pi\boldsymbol{\Omega})$$

The Inner Optimization Problem

The inner optimization problem consists of finding the η_i that maximizes the individual l_i (for a given θ). Gradient based optimization methods need accurate gradients. The k^{th} component of the gradient of the log-likelihood wrt η_i

$$\frac{dl_i}{d\eta_{ik}} = -\frac{1}{2}\sum_{j=1}^{n_i}\left(2\boldsymbol{\epsilon}_{ij}^T\mathbf{R}_{ij}^{-1}\frac{d\boldsymbol{\epsilon}_{ij}}{d\eta_{ik}} - \boldsymbol{\epsilon}_{ij}^T\mathbf{R}_{ij}^{-1}\frac{d\mathbf{R}_{ij}}{d\eta_{ik}}\mathbf{R}_{ij}^{-1}\boldsymbol{\epsilon}_{ij} + \mathrm{tr}\left[\mathbf{R}_{ij}^{-1}\frac{d\mathbf{R}_{ij}}{d\eta_{ik}}\right]\right) - \boldsymbol{\eta}_i^T\boldsymbol{\Omega}^{-1}\frac{d\boldsymbol{\eta}_i}{d\eta_{ik}}$$

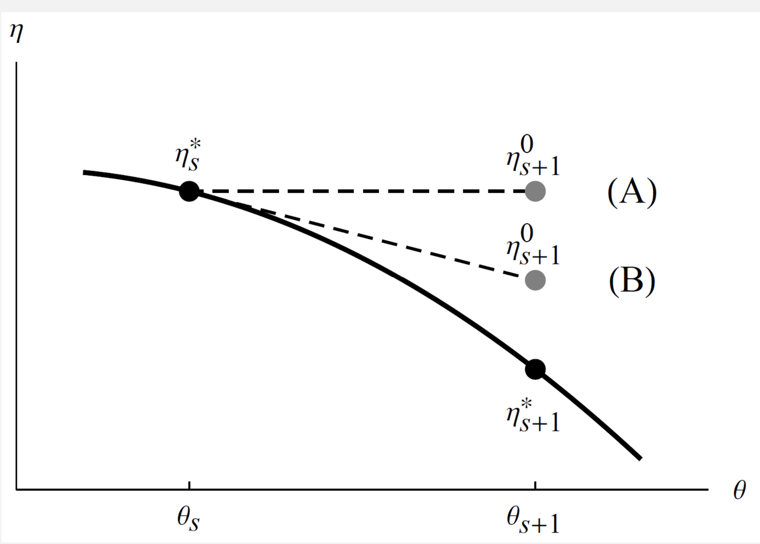
where

$$\frac{d\boldsymbol{\epsilon}_{ij}}{d\eta_{ik}} = \frac{d(\mathbf{d}_{ij} - \hat{\mathbf{y}}_{ij})}{d\eta_{ik}} = -\left(\frac{\partial \mathbf{h}}{\partial \eta_{ik}} + \frac{\partial \mathbf{h}}{\partial \mathbf{x}_{ij}}\frac{d\mathbf{x}_{ij}}{d\eta_{ik}}\right) \quad \text{and} \quad \frac{d\mathbf{R}_{ij}}{d\eta_{ik}} = \frac{\partial \mathbf{R}_{ij}}{\partial \eta_{ik}} + \frac{\partial \mathbf{R}_{ij}}{\partial \mathbf{x}_{ij}}\frac{d\mathbf{x}_{ij}}{d\eta_{ik}}$$

The *sensitivity* differential equations wrt η_{ik}

$$\frac{d}{dt}\left(\frac{d\mathbf{x}_i}{d\eta_{ik}}\right) = \frac{\partial \mathbf{f}}{\partial \eta_{ik}} + \frac{\partial \mathbf{f}}{\partial \mathbf{x}_i}\left(\frac{d\mathbf{x}_i}{d\eta_{ik}}\right) \qquad \left(\frac{d\mathbf{x}_i}{d\eta_{ik}}\right)(t_0) = \frac{\partial \mathbf{x}_{0i}}{\partial \eta_{ik}}$$

Starting Values for Random Parameters



Using that $\eta_i^* = \eta_i^*(\theta)$ is a function of θ and that we have $\frac{d\eta_i^*}{d\theta}$ gives improved starting values of the inner optimization problem

$$\eta_{s+1}^0 = \eta_s^* + \frac{d\eta_s^*}{d\theta}(\theta_{s+1} - \theta_s)$$

Acknowledgements

This project has in part been supported by the Swedish Foundation for Strategic Research, which is gratefully acknowledged.

The Outer Optimization Problem

The outer optimization problem consists of finding the θ that maximizes the log-likelihood. The m^{th} component of the gradient of the log-likelihood wrt θ

$$\frac{d\log L_F}{d\theta_m} = \sum_{i=1}^{N}\left(\frac{dl_i(\boldsymbol{\eta}_i^*)}{d\theta_m} - \frac{1}{2}\mathrm{tr}\left[\mathbf{H}_i^{-1}(\boldsymbol{\eta}_i^*)\frac{d\mathbf{H}_i(\boldsymbol{\eta}_i^*)}{d\theta_m}\right]\right)$$

where the total derivatives of l_i and H_i wrt θ can be expressed in terms of solutions to the sensitivity differential equations, e.g.,

$$\frac{dl_i(\boldsymbol{\eta}_i^*)}{d\theta_m} = \left.\frac{dl_i(\boldsymbol{\eta}_i)}{d\theta_m}\right|_{\boldsymbol{\eta}_i = \boldsymbol{\eta}_i^*(\boldsymbol{\theta})} = \left[-\frac{1}{2}\sum_{j=1}^{n_i}\left(2\boldsymbol{\epsilon}_{ij}^T\mathbf{R}_{ij}^{-1}\frac{d\boldsymbol{\epsilon}_{ij}}{d\theta_m} - \boldsymbol{\epsilon}_{ij}^T\mathbf{R}_{ij}^{-1}\frac{d\mathbf{R}_{ij}}{d\theta_m}\mathbf{R}_{ij}^{-1}\boldsymbol{\epsilon}_{ij} + \mathrm{tr}\left[\mathbf{R}_{ij}^{-1}\frac{d\mathbf{R}_{ij}}{d\theta_m}\right]\right) + \frac{1}{2}\boldsymbol{\eta}_i\boldsymbol{\Omega}^{-1}\frac{d\boldsymbol{\Omega}}{d\theta_m}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}_i - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}\frac{d\boldsymbol{\Omega}}{d\theta_m}\right]\right]_{\boldsymbol{\eta}_i = \boldsymbol{\eta}_i^*(\boldsymbol{\theta})}$$

$$\frac{d\boldsymbol{\epsilon}_{ij}^*}{d\theta_m} = \left.\frac{d\boldsymbol{\epsilon}_{ij}}{d\theta_m}\right|_{\boldsymbol{\eta}_i = \boldsymbol{\eta}_i^*(\boldsymbol{\theta})} + \left.\frac{d\boldsymbol{\epsilon}_{ij}}{d\boldsymbol{\eta}_i}\right|_{\boldsymbol{\eta}_i = \boldsymbol{\eta}_i^*(\boldsymbol{\theta})}\frac{d\boldsymbol{\eta}_i^*}{d\theta_m} \quad \text{where} \quad \frac{d\boldsymbol{\epsilon}_{ij}}{d\theta_m} = \frac{d(\mathbf{d}_{ij} - \hat{\mathbf{y}}_{ij})}{d\theta_m} = -\left(\frac{\partial \mathbf{h}}{\partial \theta_m} + \frac{\partial \mathbf{h}}{\partial \mathbf{x}_{ij}}\frac{d\mathbf{x}_{ij}}{d\theta_m}\right)$$

The sensitivity differential equations wrt θ_m

$$\frac{d}{dt}\left(\frac{d\mathbf{x}_i}{d\theta_m}\right) = \frac{\partial \mathbf{f}}{\partial \theta_m} + \frac{\partial \mathbf{f}}{\partial \mathbf{x}_i}\left(\frac{d\mathbf{x}_i}{d\theta_m}\right) \qquad \left(\frac{d\mathbf{x}_i}{d\theta_m}\right)(t_0) = \frac{\partial \mathbf{x}_{0i}}{\partial \theta_m}$$

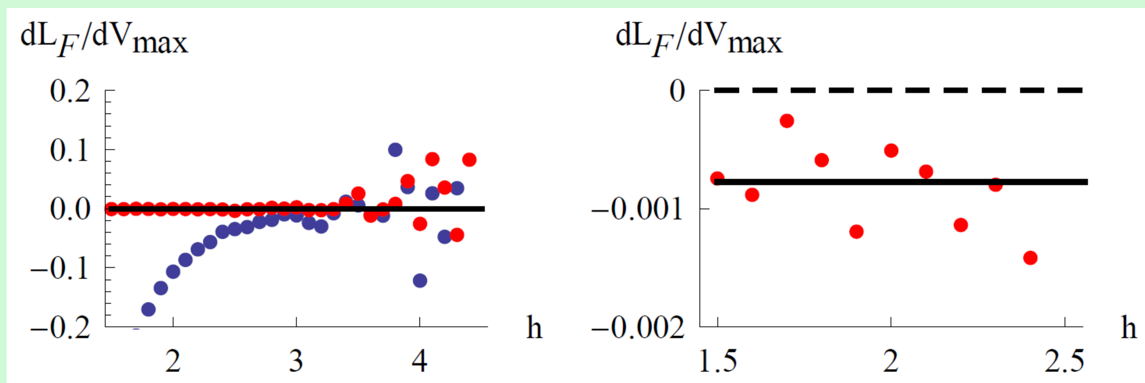
How to find $\frac{d\eta_i^*}{d\theta}$?

$$\left.\frac{dl_i}{d\eta_i}\right|_{\eta_i = \eta_i^*(\theta)} = 0 \;\Rightarrow\; \frac{d}{d\theta}\left(\left.\frac{dl_i}{d\eta_i}\right|_{\eta_i = \eta_i^*(\theta)}\right) = 0 \;\Rightarrow\; \frac{d\eta_i^*}{d\theta} = -\left(\left.\frac{d^2 l_i}{d\eta_i^2}\right|_{\eta_i = \eta_i^*(\theta)}\right)^{-1}\left.\frac{d^2 l_i}{d\eta_i d\theta}\right|_{\eta_i = \eta_i^*(\theta)}$$

Second order sensitivities are also required: $\frac{d^2 x_i}{d\eta_{ik} d\theta_m}$ and $\frac{d^2 x_i}{d\eta_{ik} d\eta_{il}}$.

Precision, Accuracy, and Performance

Two different levels of magnification of an element of the log-likelihood gradient as a function of the finite difference step, h.



Forward diff (blue)

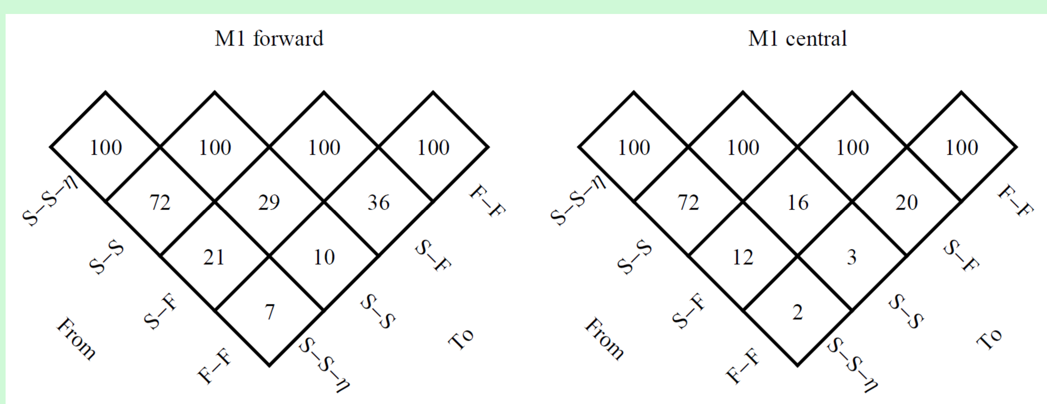
$$\frac{\log L_F(\theta_m(1+10^{-h})) - \log L_F(\theta_m)}{\theta_m 10^{-h}}$$

Central diff (red)

$$\frac{\log L_F(\theta_m(1+10^{-h})) - \log L_F(\theta_m(1-10^{-h}))}{2\theta_m 10^{-h}}$$

Sensitivity based (solid black line)

Benchmarking – relative estimation times



Model M1: 2-compartment, nonlinear elimination

S-F- η: **S**ensitivities (inner), **F**inite differences (outer), improved $\boldsymbol{\eta}$ starting values

Example: F-F (central diff) to S-S-η gives 50-fold decreased computational time

Highlights

- Robust computation of gradients
- Methodology applies to both individual and population log-likelihoods
- Improves computational speed compared to finite differences

References

[1] Wang Y. Derivation of various NONMEM estimation methods. J of Pharmacokin Pharmacodyn (2007) 34(5): 575-593.
[2] Johansson ÅM, Ueckert S, Plan EL, Hooker AC, Karlsson MO. Evaluation of bias, precision, robustness and runtime for estimation methods in NONMEM 7. J of Pharmacokin Pharmacodyn (2014) 41(3):223-238.
[3] Almquist J, Leander J, Jirstrand M. Using sensitivity equations for computing gradients of the FOCE and FOCEI approximations to the population likelihood. J of Pharmacokin Pharmacodyn (2015) 42(3):191-209.