A Fast Stochastic Approximation of the EM Algorithm with Applications to Pharmacology

Belhal Karimi

Collaboration with **Marc Lavielle** and **Eric Moulines**

LIXOFT

Maximum Likelihood Estimation (MLE)

- The MLE problem is, given a model $g(Y, \theta)$ and some actual data Y, find the parameter θ which makes the data most likely:

$$\theta^{ML} := \arg\max_{\theta} \; g(Y, \theta)$$

- This problem is an **optimization problem**, which we could use any imaginable tool to solve
- In practice, it's often **hard** to get expressions for the **derivatives** needed by **gradient** methods
- **Expectation-Maximization (EM)** method is one popular and powerful way of proceeding, but not the only way. **It takes advantage of the latent data to complete the observations.**

EM Algorithm

[Dempster, Laird and Rubin, 1977]

- **E-step:** Given $\theta^{(k-1)}$ compute the surrogate quantity

$$Q(\theta, \theta^{(k-1)}) = \mathbb{E}_{p(z|y,\theta^{(k-1)})} \left[\log f(z, y, \theta)\right]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{p(z_i|y_i,\theta^{(k-1)})} \left[\log f\left(z_i, y_i; \theta\right)\right]$$

- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(k-1)})$$

Expectation

Conditional Distribution

Large Sum

Pharmacology and Mixed Effects Modeling

- **Pharmacokinetics:** Evolution of drug in the body
  - Outcome can be plasma drug concentration
- **Pharmacodynamics:** Reaction of the body to a drug
  - Evolution of number of seizures after treatment

Settings and Notations

- **Population approach:** Consider n individuals. Vector of measurements for each individual $y_i = (y_{ij}, 1 \leq j \leq n_i)$
- **Latent Data Model:** The latent variables are called 'individual parameters' ψ_i
- **Parametrized Hierarchical model:**

$$y_i \sim p(y_i|\psi_i, \theta) \qquad \psi_i \sim p(\psi_i, \theta)$$

- **Mixed Effects Model:** The individual parameters can be decomposed as follows $\psi_i = G(\beta, \eta_i)$

β : Population parameter η_i : Random effects $\eta_i \sim \mathcal{N}(0, \Omega)$

Continuous data model

- Continuous, nonlinear and mixed effects model

$$y_{ij} = f(t_{ij}; \psi_i) + \epsilon_{ij}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \qquad \psi_i = \beta + \eta_i \Rightarrow \psi_i \sim \mathcal{N}(\beta, \Omega)$$
$$\theta = (\beta, \Omega, \sigma)$$

Non Continuous data model

- No analytical relationship between observations and individual parameters
- **Repeated time-to-event models**

$$\mathbb{P}(T_{ij} > t | T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^{t} h(u,\psi_i)\mathrm{d}u}$$

EM Algorithm

Updates

- **E-step:** Given $\theta^{(k-1)}$:

$$Q(\theta, \theta^{(k-1)}) = \sum_{i=1}^{n} \mathbb{E}_{p(\psi_i | y_i, \theta^{(k-1)})} [\log f(\psi_i, y_i, \theta)]$$

- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(k-1)})$$

SAEM Algorithm

Updates [Delyon+, 1999]

- **E-step:** Given $\theta^{(k-1)}$:
  - **Simulation Step:** $\psi_i^{(k)} \sim p(\psi_i | y_i, \theta^{(k-1)})$
  - **Stochastic Approximation of** $Q(\theta, \theta^{(k-1)})$ **:**

$$Q^{(k)}(\theta) = Q^{(k-1)}(\theta) + \gamma_k \left(\sum_{i=1}^{n} \log f(\psi_i^{(k)}; y_i, \theta) - Q^{(k-1)}(\theta) \right)$$

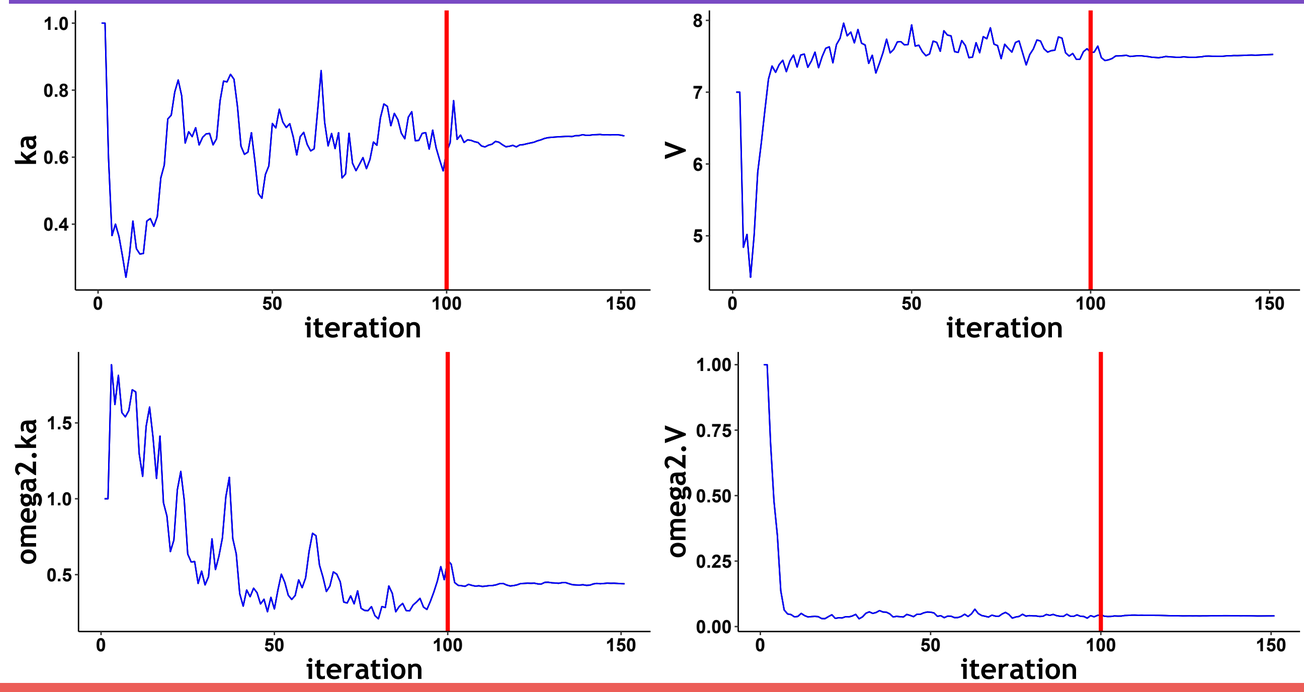
- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q^{(k)}(\theta)$$

Speed up the sampling procedure to speed up the MLE algorithm

How is it behaving?

- For a simple pharmacokinetics (PK) model
- In practice the stepwise is equal to 1 during the first K1 iterations then decreases in $1/k^{\alpha}$



Posterior Sampling Procedures

Metropolis-Hastings (MH)

- Sampling a candidate from a proposal

$$\psi^c \sim q(\psi|\psi^{(t-1)})$$

- Compute MH ratio

$$\alpha(\psi_i^{(t-1)}, \psi_i^c) = \frac{p(\psi_i^c|y_i)}{p(\psi_i^{(t-1)}|y_i)} \frac{q_i(\psi_i^{(t-1)}|\psi_i^c)}{q_i(\psi_i^c|\psi_i^{(t-1)})}$$

- Accept or reject with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(t-1)}))$

Can Be SLOW

Metropolis Adjusted Langevin [Roberts+, 1998]

- Using the gradient of the target distribution

$$\psi_i^c \sim \mathcal{N}(\psi_i^{(t)} - \gamma_t \nabla \log \pi(\psi_i^{(t)}), 2\gamma_t)$$

- Special case of RWM [Ma+, 2015] with covariance matrix that is diagonal and isotropic

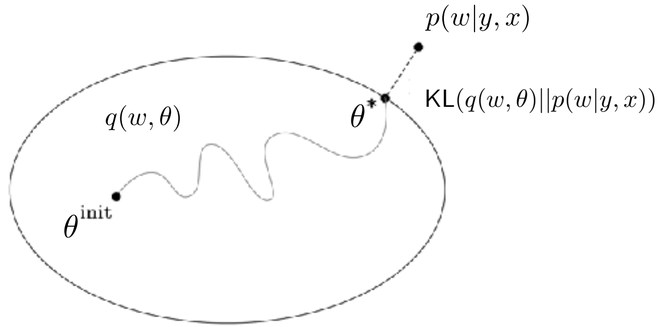
COSTLY and NEED TUNING

The No-U-Turn Samplers [Hoffman+, 2011]

- Adaptive and automated Hamiltonian Monte Carlo
- Using Hamiltonian dynamics
- State-of-the-art sampler

VERY COSTLY

The Variational MCMC [de Freitas, 2013]



Need to design **efficient** and easy-to-**implement** proposal based on the **covariance structure** of the latent variables

Efficient Metropolis-Hastings Procedure

[Karimi and Lavielle, BAYSM 2018]

- **Maximum A Posteriori**

$$\hat{\psi}_i = \arg\max_{\psi_i} p(\psi_i|y_i,\theta) = \arg\max_{\psi_i} p(y_i|\psi_i,\theta)p(\psi_i,\theta)$$

- **General Data Models**
  - Compute the Laplace Approximation of the incomplete likelihood

$$g(y_i,\theta) = \int e^{\log f(y_i,\psi_i,\theta)} \mathrm{d}\psi_i$$

  - Taylor expansion of the complete log likelihood around the MAP

$$\log p(\hat{\psi}_i|y_i,\theta) \approx -\frac{p}{2}\log 2\pi - \frac{1}{2}\log\left(\left|-\nabla^2 \log p(y_i,\hat{\psi}_i,\theta)\right|\right)$$

Proposal

$$\mathcal{N}(\mu_i,\Gamma_i) \qquad \mu_i = \hat{\psi}_i$$

$$\Gamma_i = -\nabla^2 \log p(y_i,\hat{\psi}_i,\theta)^{-1} = -\left(\nabla^2 \log p(y_i|\hat{\psi}_i,\theta) + \Omega^{-1}\right)^{-1}$$

- **Continuous Data Models**
  - Taylor expansion of the structural model around the MAP

$$f(\psi_i) \approx f(\hat{\psi}_i) + \mathsf{J}_f^{\psi_i}(\hat{\psi}_i)(\psi_i - \hat{\psi}_i)$$

  - Relation between observation and individual parameter is now linear and the posterior is a tractable Normal distribution

Proposal

$$\mathcal{N}(\mu_i,\Gamma_i) \qquad \mu_i = \hat{\psi}_i$$

$$\Gamma_i = \left(\frac{\mathsf{J}_f^{\psi_i}(\hat{\psi}_i)^\top \mathsf{J}_f^{\psi_i}(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right)^{-1}$$

f-SAEM Algorithm

[Karimi+, CSDA 2019]

- **E-step:** Given $\theta^{(k-1)}$:
  - **Simulation Step:** $\psi_i^{(k)} \sim p(\psi_i|y_i, \theta^{(k-1)})$
  - **Stochastic Approximation of** $Q(\theta, \theta^{(k-1)})$ **:**

$$Q^{(k)}(\theta) = Q^{(k-1)}(\theta) + \gamma_k \left(\sum_{i=1}^{n} \log f(\psi_i^{(k)}; y_i, \theta) - Q^{(k-1)}(\theta) \right)$$

- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q^{(k)}(\theta)$$

nlme-MH Algorithm

Algorithm 5.3 The nlme-IMH algorithm

Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution ξ_i .

- Compute the MAP estimate:

$$\hat{\psi}_i = \arg\max_{\psi_i \in \mathbb{R}^p} p_i(\psi_i|y_i, \theta) \ .$$

- Compute the covariance matrix Γ_i using the corresponding proposal.

Iteration t: Given the current state of the chain $\psi_i^{(t-1)}$:

1. Sample a candidate ψ_i^c from a the independent proposal $\mathcal{N}(\hat{\psi}_i, \Gamma_i)$ denoted $q_i(\cdot|\hat{\psi}_i)$.
2. Compute the MH ratio:

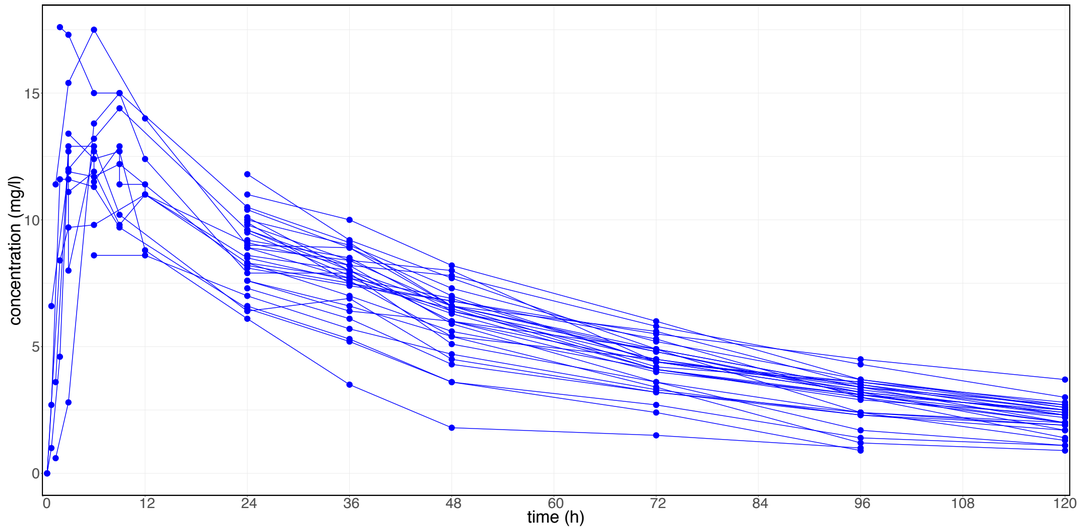
$$\alpha(\psi_i^{(t-1)}, \psi_i^c) = \frac{p_i(\psi_i^c|y_i, \theta)}{p_i(\psi_i^{(t-1)}|y_i, \theta)} \frac{q_i(\hat{\psi}_i|\psi_i^c)}{q_i(\psi_i^c|\hat{\psi}_i)} \ .$$

3. Set $\psi_i^{(t)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(t-1)}))$ (otherwise, keep $\psi_i^{(t)} = \psi_i^{(t-1)}$).

Numerical Applications

Warfarin Data

- 32 healthy volunteers received a 1.5 mg/kg single oral dose of warfarin, an anticoagulant normally used in the prevention of thrombosis



- **Goal:** fit a PK model on this dataset
  - Compute population parameters
  - Obtain a fitted model for predictive tasks

PK Model

- Continuous nonlinear mixed effects model

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

- Lognormally distributed individual parameters

$$\log(\psi_i) \sim \mathcal{N}(\log(\psi_{\text{pop}}), \omega_\psi^2)$$

- One-compartment PK model for oral administration, assuming first-order absorption and linear elimination processes:

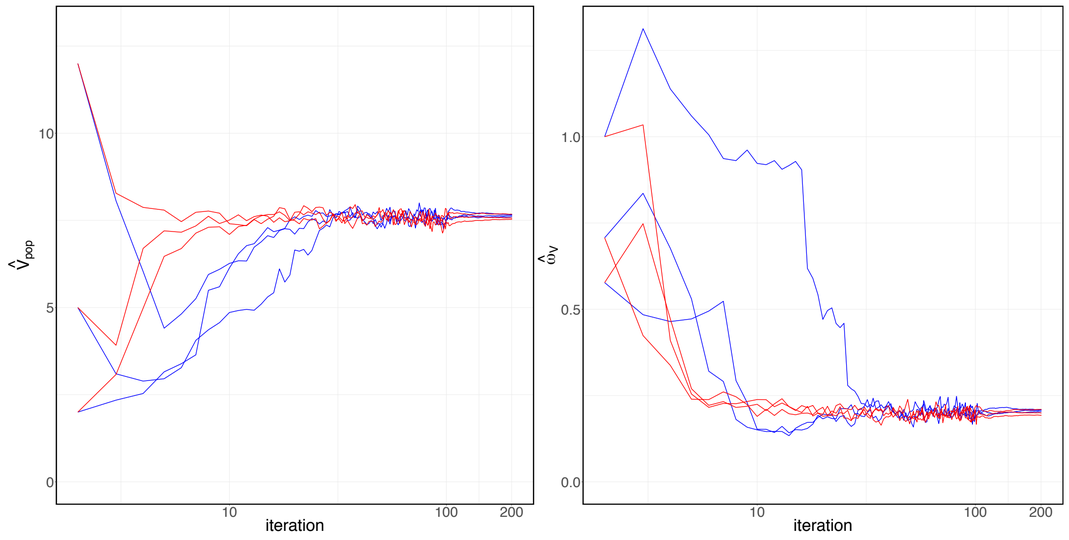
$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(\mathrm{e}^{-ka\, t} - \mathrm{e}^{-k\, t})$$

$$\psi_i = (ka_i, V_i, k_i) \qquad \theta = (ka_{\text{pop}}, \cdots, \omega_{ka}, \cdots, \sigma)$$

Numerical Applications

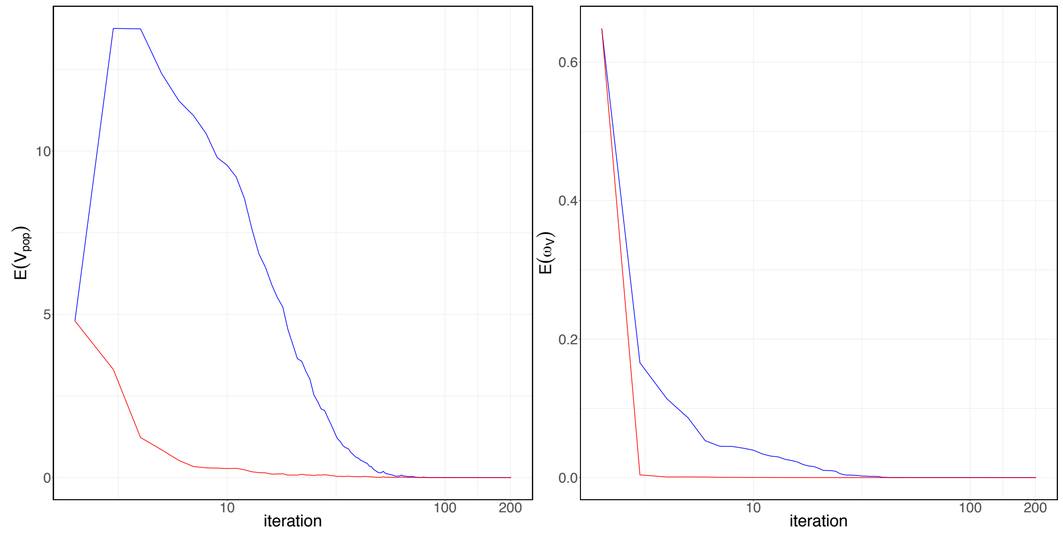
Warfarin Data: MLE Convergence

- **Single run MLE**
  - Three different initialization
  - K1 = 100, K2 = 100
  - 6 MCMC transitions per iteration



- **Monte Carlo Study**
  - M = 50 synthetic datasets
  - M runs of K iterations to obtain vector of ML estimates

$$E_k(\ell) = \frac{1}{M} \sum_{m=1}^{M} \left(\theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell) \right)^2$$



- Estimation of V_{pop} and ω_V
- Reference (RWM) in Blue and f-SAEM in Red

Numerical Applications

Time-to-event Data Model

- **Time-to-event Data Model**

$$\mathbb{P}(T_{ij} > t | T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^{t} h(u,\psi_i)\mathrm{d}u}$$

- **Weibull model** for time-to-event data. Hazard function is defined as

$$h(t,\psi_i) = \frac{\beta_i}{\lambda_i}\left(\frac{t}{\lambda_i}\right)^{\beta_i - 1}$$

- Two parameters are independent and log normally distributed

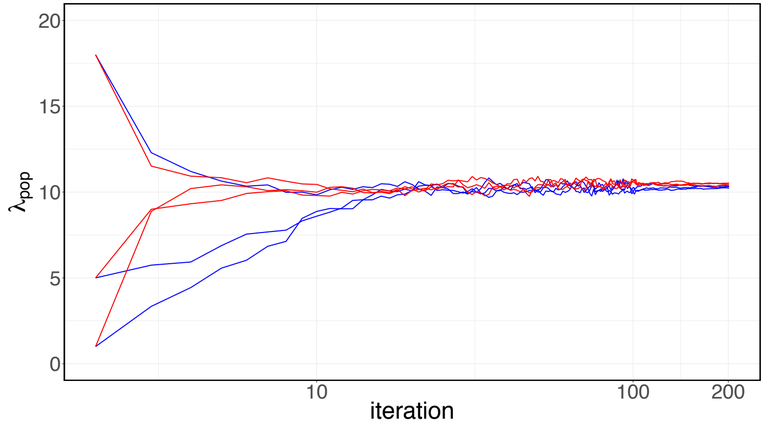
$$\log(\lambda_i) \sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2)$$
$$\log(\beta_i) \sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2)$$

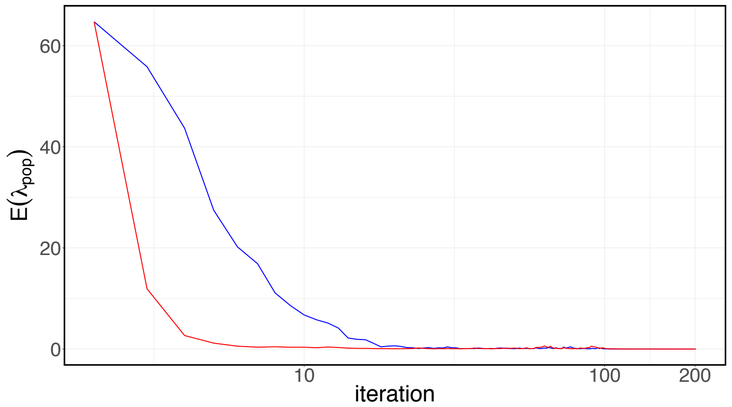
- Vector of parameters to estimate

$$\theta = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda, \omega_\beta)$$

Experiments

- Synthetic data: n= 100 individuals and right censoring time of 20
- Posterior sampling: comparison between reference RWM in blue and nlme-IMH in red





iSAEM Algorithm

Updates for exponential family

- **E-step:** Given $\theta^{(k-1)}$:
  - **Simulation Step:** $\psi_i^{(k)} \sim p(\psi_i|y_i, \theta^{(k-1)})$
  - **Stochastic Approximation of** $\bar{s}(y, \bar{\theta}(\hat{s}_{k-1}))$**:**

$$\hat{s}_i^{(k)} = \hat{s}_i^{(k-1)} + \gamma_k \left(S(\psi_i^{(k)}, y_i) - \hat{s}_i^{(k-1)} \right)$$

- **M-step:** Maximization function

$$\theta^{(k)} = \bar{\theta}(\hat{s}_k) \qquad \hat{s}_k = \left(\hat{s}_1^{(k)}, \cdots, \hat{s}_n^{(k)} \right)$$

$$\bar{\theta}(s) := \arg\max_{\theta \in \Theta} \langle s | \phi(\theta) \rangle - \psi(\theta) - R(\theta)$$

Incremental Updates

- **Simulation step:** Sample latent variables only for a mini batch of indices sampled uniformly

$$I_k \sim \{A \subset [1, n], \mathrm{card}(A) = p\}$$

- **Update:** Minibatch of sufficient statistics component are updated. The others remain unchanged

$$\hat{s}_i^{(k)} = \begin{cases} \hat{s}_i^{(k-1)} + \gamma_k (S_i(\psi_i^{(k)}, y_i) - \hat{s}_i^{(k-1)}) & \text{if } i \in I_k. \\ \hat{s}_i^{(k-1)} & \text{otherwise.} \end{cases}$$

- **Maximization** remains unchanged

Numerical Applications

Gaussian Mixture Models (GMM)

- Fit a GMM model to a set of n observations
- Each of M components with unit variance
- The complete log likelihood reads:

$$\log f\left(z_i, y_i; \boldsymbol{\theta}\right) = \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\left[\log\left(\omega_m\right) - \mu_m^2/2\right] + \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\mu_m y_i + \text{ constant}$$

$$\theta := (\omega, \mu)$$

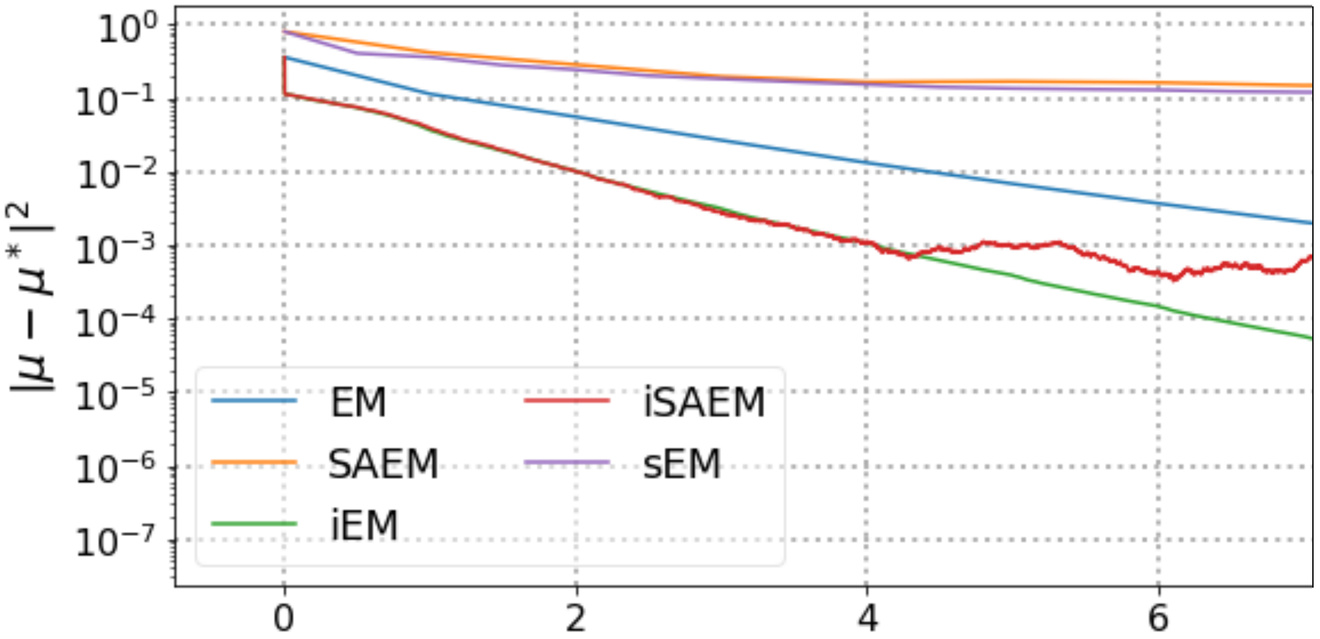
$$\omega = \{\omega_m\}_{m=1}^{M-1}$$

$$\mu = \{\mu_m\}_{m=1}^{M}$$

- Penalization used: $\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{m=1}^{M}\mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$

Experiments

- Numerical: GMM with M=2 and $\mu_1 = -\mu_2 = 0.5$
- **Fixed sample size:** size $n = 10^3$ and run to get μ^*
  Stepsize for sEM $\gamma_k = 3/(k+10)$
  Stepsize for iSAEM $\gamma_k = 1/k^{0.6}$
- Compare to iEM, sEM and Batch EM



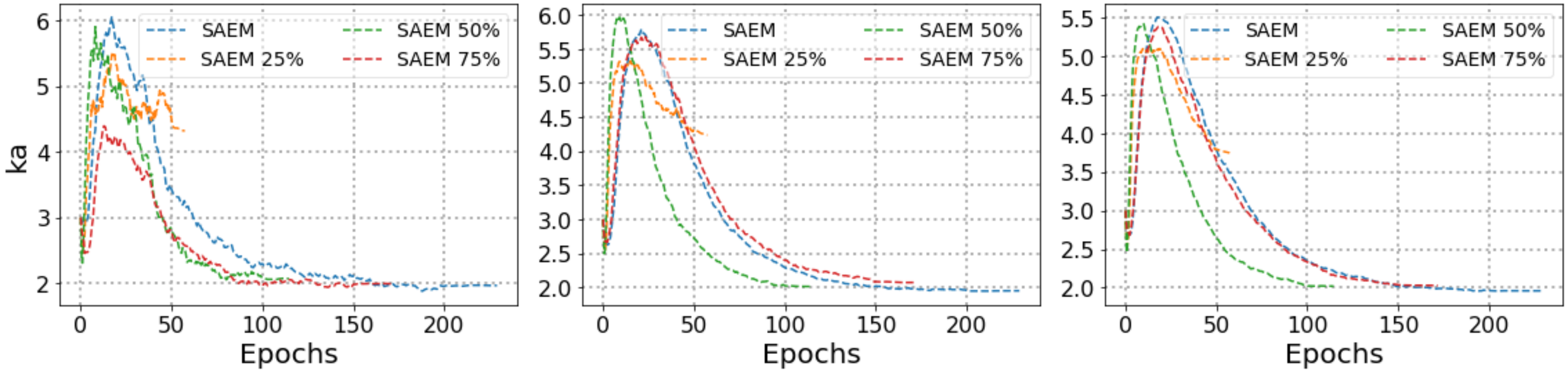
Numerical Applications

One-compartment PK Model

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(\mathrm{e}^{-ka\, t} - \mathrm{e}^{-k\, t}) \qquad \log(\psi_i) \sim \mathcal{N}(\log(\psi_{\text{pop}}), \omega_\psi^2) \qquad \theta = (ka_{\text{pop}}, \cdots, \omega_{ka}, \cdots, \sigma)$$

Synthetic data

- Simulate observations for $n = 10^3$ individuals with $n_i = 5$
- Stepsizes set to $\gamma_k = 1$ for K1 = 200 and $\gamma_k = 1/k$ for K2 = 50
- Run for three different size of MC batch size: 1, 10 and 20

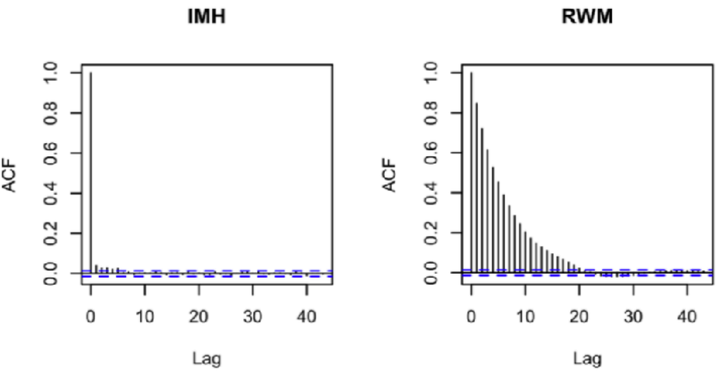


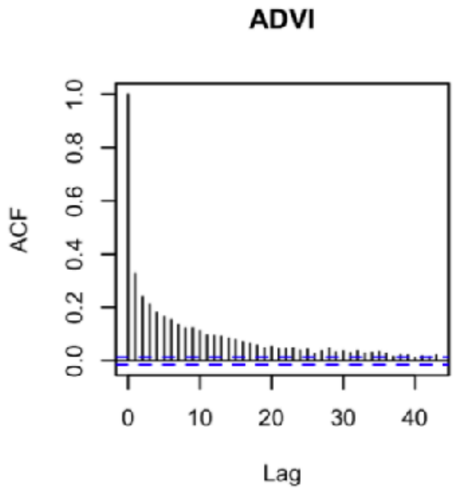
Thank You !

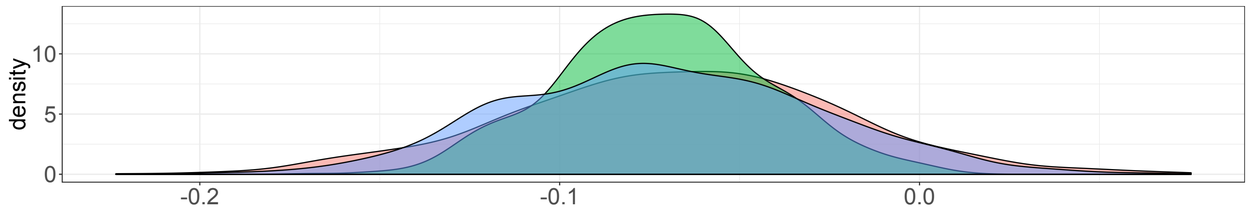
Appendix

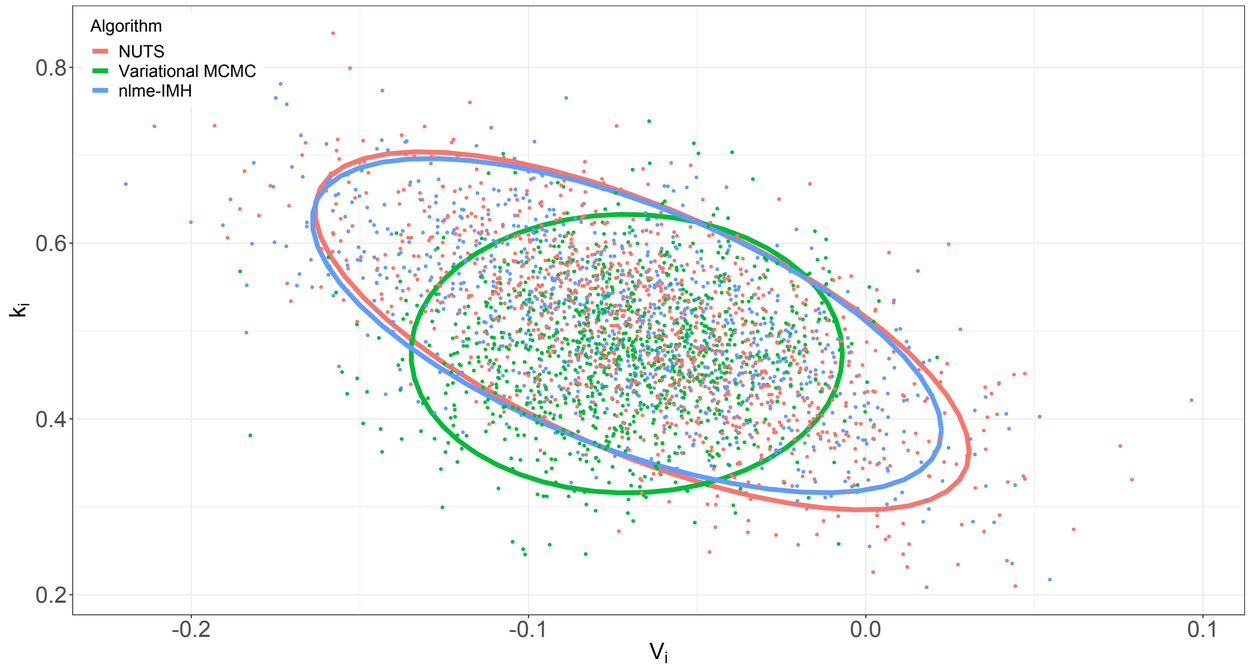
Numerical Applications

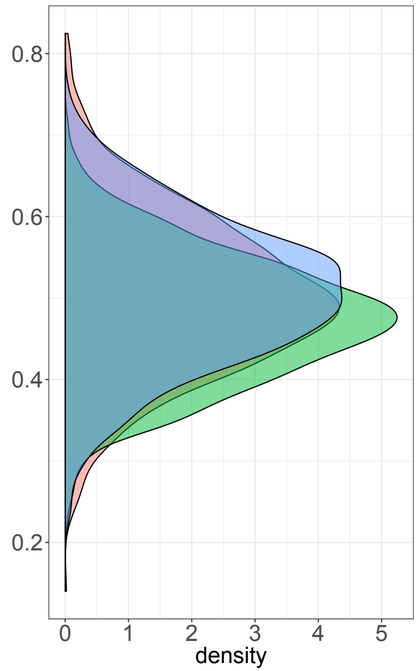
Warfarin Data: Posterior Sampling Experiment (VARIATIONAL MCMC/ ADVI)







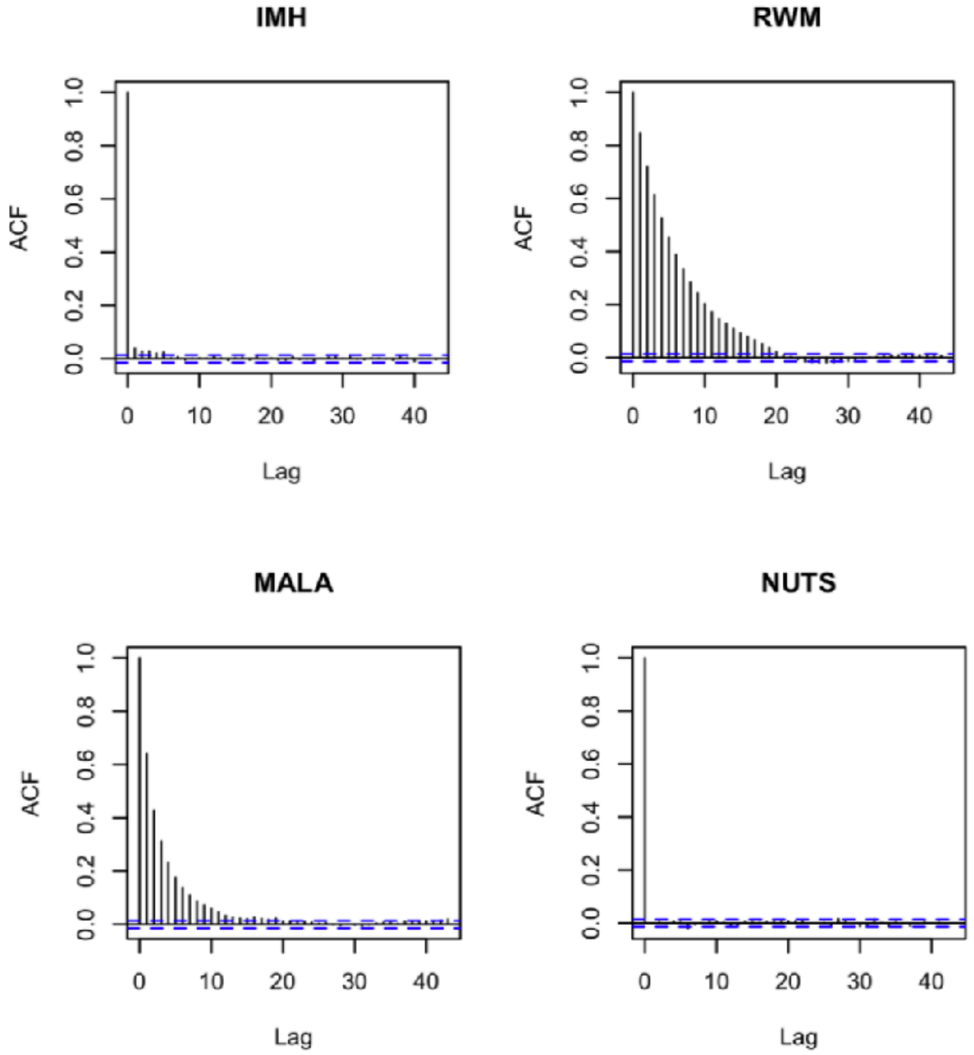




Numerical Applications

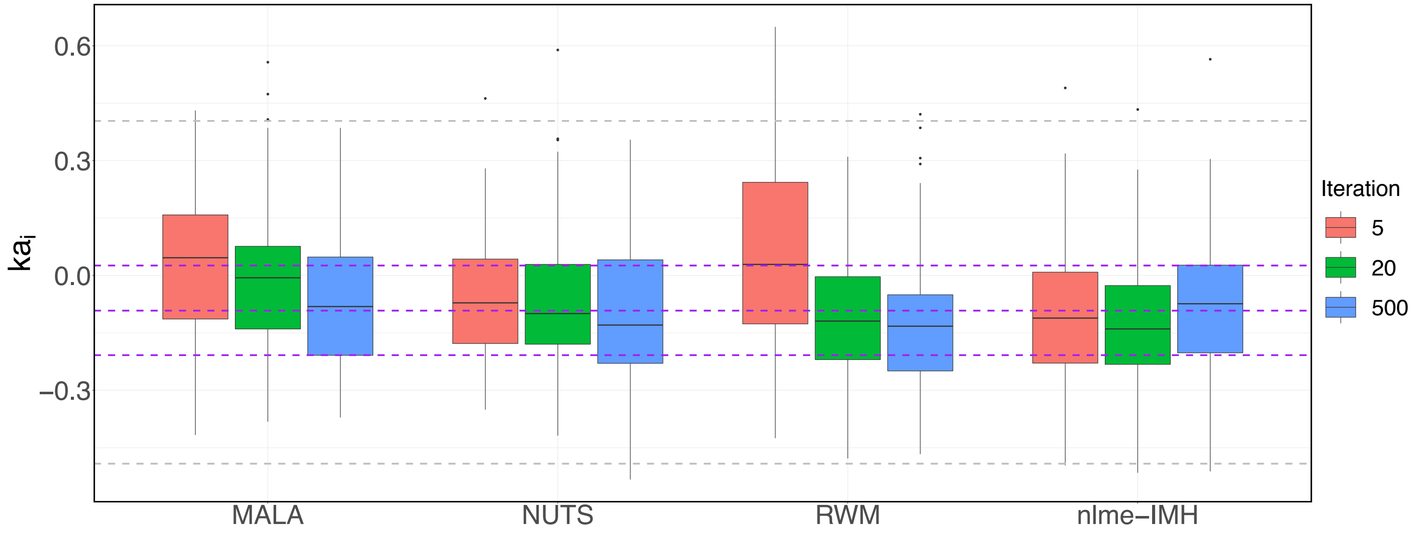
- Posterior sampling of $k_i|y_i$ for a given individual and at θ close to θ^{ML}

Warfarin Data: Posterior Sampling Experiment



	MSJD			ESS		
	ka_i	V_i	k_i	ka_i	V_i	k_i
RWM	0.009	0.002	0.006	1728	3414	3784
nlme-IMH	0.061	0.004	0.018	13694	14907	19976
MALA	0.024	0.002	0.006	3458	3786	3688
NUTS	0.063	0.004	0.018	18684	19327	19083

First ~500 iterations



Numerical Applications

Time-to-event Data Model

- **Time-to-event Data Model**

$$\mathbb{P}(T_{ij} > t | T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^{t} h(u,\psi_i)\mathrm{d}u}$$

- **Weibull model** for time-to-event data. Hazard function is defined as

$$h(t,\psi_i) = \frac{\beta_i}{\lambda_i}\left(\frac{t}{\lambda_i}\right)^{\beta_i - 1}$$

- Two parameters are independent and log normally distributed

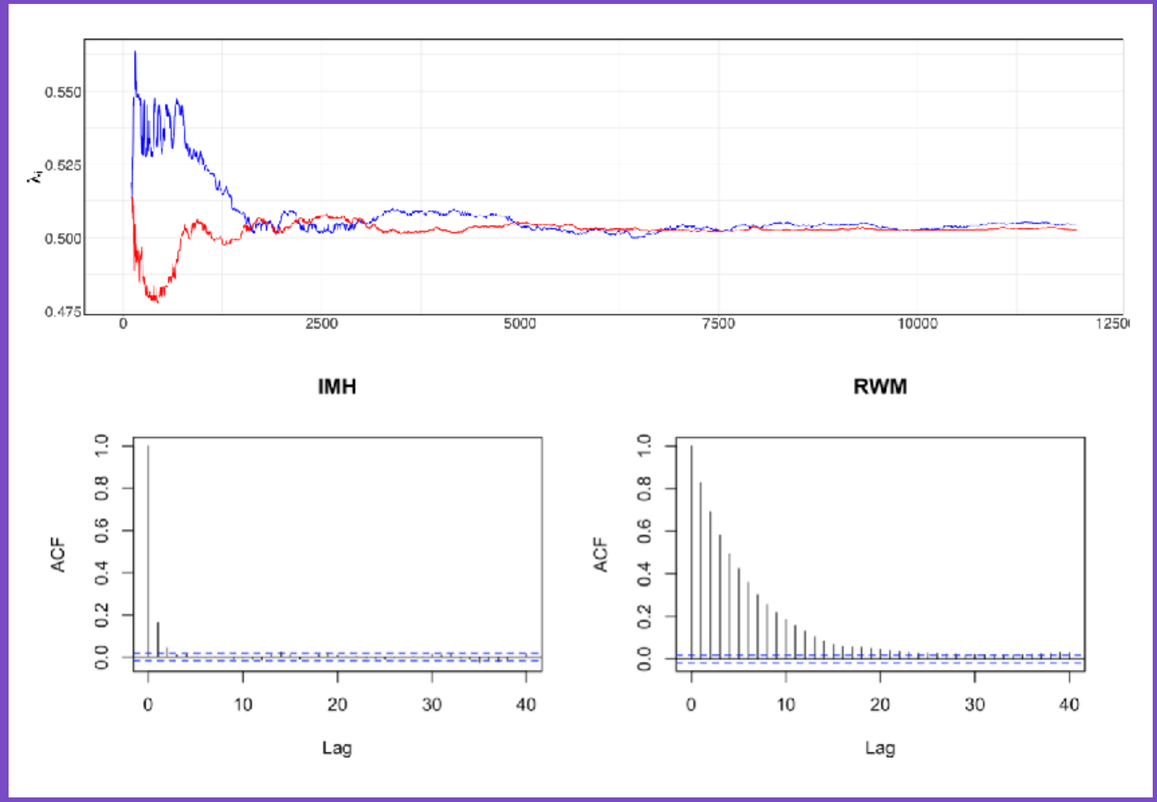
$$\log(\lambda_i) \sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2)$$
$$\log(\beta_i) \sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2)$$

- Vector of parameters to estimate

$$\theta = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda, \omega_\beta)$$

Experiments

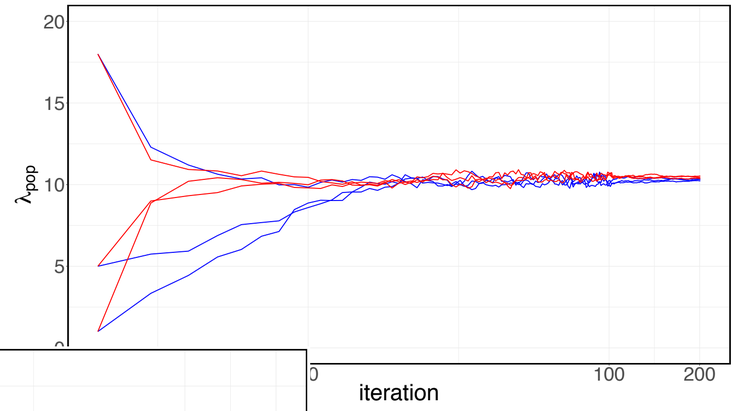
- Synthetic data: n= 100 individuals and right censoring time of 20
- Posterior sampling: comparison between reference RWM in blue and nlme-IMH in red

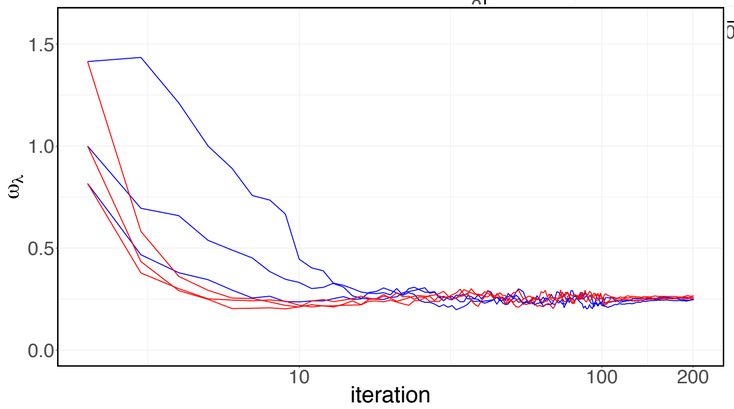


Numerical Applications

Time-to-event Data Model

- **Single run MLE**
  - Three different initialization





- Estimation of λ_{pop} and ω_λ
- Reference (RWM) in Blue and f-SAEM in Red

- **Monte Carlo Study**
  - M synthetic datasets
  - M runs of K iterations to obtain vector of ML estimates

$$E_k(\ell) = \frac{1}{M} \sum_{m=1}^{M} \left(\theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell) \right)^2$$