Some Accelerations of MLE Algorithm

COMPSTATS 2018

Belhal Karimi, Marc Lavielle and Eric Moulines

Aug. 29^{th}, 2018

Settings and Notations

- **Population approach**. Consider N individuals.
  $y_i = (y_{ij}, 1 \leq j \leq n_i)$ vector of n_i measurements for individual i and c_i individual covariates.
- **Incomplete data** Individual parameters ψ_i are latent.
- **Parametrized hierarchical model**. The distribution of y_i depends on the latent variable ψ_i

$$
\begin{aligned}
y_i &\sim p(y_i|\psi_i, \theta) \\
\psi_i &\sim p(\psi_i|c_i, \theta) \\
c_i &\sim p(c_i, \theta)
\end{aligned}
\tag{1}
$$

- **Mixed Effects Model**. The individual parameters are decomposed as follows:

$$
\psi_i = g(\beta, c_i, \eta_i) \tag{2}
$$

  where β is the population parameter (fixed effect) and η_i is the random effect. We assume $\eta_i \sim \mathcal{N}(0, \Omega)$.

Example: Continuous data model

- Continuous, non linear and mixed effects models:

$$y_{ij} = f(t_{ij}; \psi_i) + \epsilon_{ij} \tag{3}$$

  Where:

  - The structural model $f(t_{ij}, .)$ is a non linear function of ψ_i
  - $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\sigma \in \mathbb{R}$
  - $\psi_i = \beta + \eta_i \Rightarrow \psi_i \sim \mathcal{N}(\beta, \Omega)$

- Here $\theta = (\beta, \Omega, \sigma)$
- The goal is to compute the maximum likelihood estimate

$$\theta^{ML} = \arg\max_{\theta \in \Theta} p(y, \theta) \tag{4}$$

Example: Non Continuous data model

- Here, the model for the observations of individual i is the conditional distribution of y_i given the set of individual parameters ψ_i. There is no analytical relationship between the observations and the individual parameters
- For repeated event models, times when events occur for individual i are random times $(T_{ij}, 1 \leq j \leq n_i)$ for which conditional survival functions can be defined:

$$\mathbb{P}(T_{ij} > t | T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^{t} h(u,\psi_i)\mathrm{d}u} \tag{5}$$

- Then, we can show (see [5] for more details) that the conditional pdf of $y_i = (y_{ij}, 1 \leq n_i)$ writes

$$\mathrm{p}(y_i|\psi_i) = \exp\left\{-\int_0^{\tau_c} h(u,\psi_i)\mathrm{d}u\right\} \prod_{j=1}^{n_i-1} h(t_{ij},\psi_i) \tag{6}$$

Maximum likelihood: EM

The EM algorithm (Dempster, Laird and Rubin, 1977) is an iterative algorithm that computes MLE. At a given θ^{k-1}:

1. $Q(\theta, \theta^{k-1}) = \mathbb{E}_{p(\psi|y,\theta^{k-1})}\left[\log p(y, \psi, \theta)\right]$
2. $\theta^k = \arg\max_{\theta \in \Theta} Q^k(\theta)$

SAEM

Given θ^{k-1}, SAEM (Delyon, Lavielle, Moulines, 1999) $k-$th update consists in:

1. $\psi_i^k \sim p(\psi_i | y_i, \theta^{k-1})$ (MCMC [4])
2. $Q^k(\theta) = Q^{k-1}(\theta) + \gamma_k (\sum_{i=1}^N \log p(y_i, \psi_i^k, \theta) - Q^{k-1}(\theta))$
3. $\theta^k = \arg\max_{\theta \in \Theta} Q^k(\theta)$

Example (PK model):

$$f(t_{ij}; \psi_i) = \frac{Dka_i}{V_i(ka_i - k_i)}(e^{-k_i t_{ij}} - e^{-ka_i t_{ij}})$$

where $\psi_i = (ka_i, V_i, k_i)$ and

$$\begin{aligned} \log(ka_i) &\sim \mathcal{N}(\log(ka), \omega_{ka}^2) \\ \log(V_i) &\sim \mathcal{N}(\log(V), \omega_V^2) \\ \log(k_i) &\sim \mathcal{N}(\log(k), \omega_k^2) \end{aligned} \tag{7}$$

Convergence behaviour: example

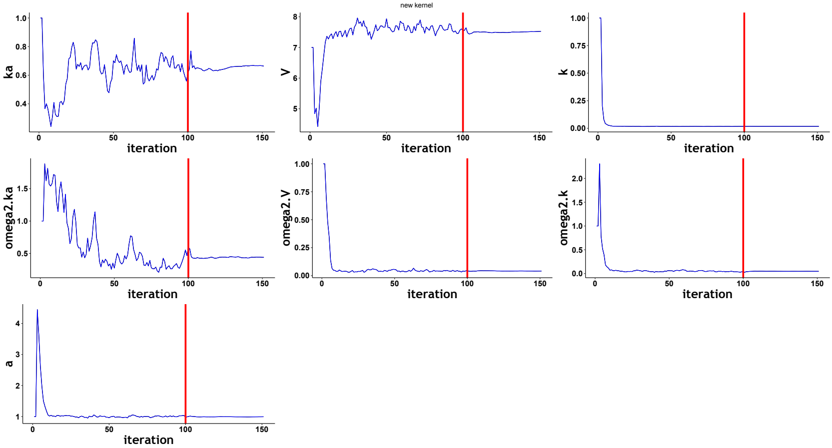

Figure 1: SAEM convergence (K1 = 100, K2 = 50)

MCMC Sampling

Metropolis Hastings:

- Random Walk Metropolis: proposals are centered in the current state with a diagonal variance-covariance matrix, with variance terms which are adaptively adjusted at each iteration in order to reach some optimal acceptance rate [1, 5]
- Attempts: SDE-based (Fox, Ma, 2015) methods using the direction of the gradient of the target distribution (tuning and heavy calculus)
- Metropolis Adjusted Langevin Algorithm (MALA) [9, 11] and its variants [7, 1]
- The Hamiltonian Monte Carlo (HMC) and its extension the "No U-Turns Sampler" [3, 2] that takes advantage of Hamiltonian dynamics to propose candidates.

Fast MCMC sampling: New proposal

In the continuous case: For a given individual i

- Compute the Maximum A Posteriori (MAP):

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

- Taylor expansion of the structural model f around this point:

$$f(\psi_i) \approx f(\hat{\psi}_i) + \nabla f(\hat{\psi}_i)(\psi_i - \hat{\psi}_i), \tag{8}$$

Proposition 1

Under this linear model, the conditional distribution of ψ_i is a Gaussian distribution with mean μ_i and variance-covariance Γ_i where

$$\begin{aligned} \mu_i &= \hat{\psi}_i, \\ \Gamma_i &= \left(\frac{\nabla f(\hat{\psi}_i)' \nabla f(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1} \right)^{-1}. \end{aligned} \tag{9}$$

Fast MCMC sampling: New proposal

In the non continuous case

- Use Laplace Approximation of the incomplete likelihood

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

- After TD of the complete log likelihood around the MAP ($\nabla \log \mathrm{p}(y_i, \hat{\psi}_i) = 0$) we obtain:

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

Proposition 2

Let (y_i, ψ_i) be a pair of random variables where ψ_i is normally distributed with variance-covariance matrix Ω. Then, the conditional distribution of ψ_i can be approximated by a Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance

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

Fast MCMC sampling: State-of-the-art

- The MALA consists in proposing a new state ψ_i^c using the gradient of the target measure at the current state $\psi_i^{(k)}$:

$$\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma_k \nabla \log \pi(\psi_i^{(k)}), 2\gamma_k), \tag{10}$$

  where $(\gamma_k)_{k>0}$ is a sequence of positive integers. It is a particular case of the RWM with a drift term [6] and a covariance matrix that is diagonal and isotropic (uniform in all directions).
- In [3], the authors proposed a novel MCMC algorithm for conditional sampling, extending the existing Hamiltonian Monte Carlo sampler. The NUTS uses a recursive algorithm to build a set of candidate samples that spans a broad range of the target distribution without requiring the user to choose how many steps it wants to execute in order to produce the candidate sample using the Hamiltonian dynamics.
  We use its implementation in rstan (R Package [10])

Fast ML Estimation: Modified SAEM

Assume that our new proposal is used at iteration k, then the simulation step of SAEM decomposes as follows:

1. compute the MAP under the current model parameter estimate θ_{k-1} for all individuals i:

$$\hat{\psi}_i^{(k)} = \arg\max_{\psi_i} \mathrm{p}(\psi_i | y_i, \theta_{k-1}). \tag{11}$$

2. Compute the covariance matrix $\Gamma_i^{(k)}$ such as:

$$\Gamma_i^{(k)} = \begin{cases} \left(\frac{\nabla f_i(\hat{\psi}_i^{(k)}) \nabla f_i(\hat{\psi}_i^{(k)})'}{\sigma^2} + \Omega^{-1} \right)^{-1} & \text{for cont. models,} \\ \left(\nabla \log \mathrm{p}(y_i | \hat{\psi}_i^{(k)})) \nabla \log \mathrm{p}(y_i | \hat{\psi}_i^{(k)})' + \Omega^{-1} \right)^{-1} & \text{otherwise.} \end{cases} \tag{12}$$

3. Run a small number of iterations of the MH algorithm with the proposal $\mathcal{N}(\hat{\psi}_i^{(k)}, \Gamma_i^{(k)})$.

Numerical Experiment: 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 [8].

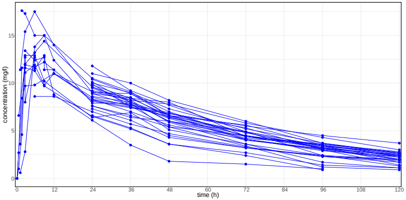

Figure 2: Warfarin concentration (mg/l) over time (h) for 32 subjects

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

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(e^{-ka\, t} - e^{-k\, t}), \tag{13}$$

MCMC convergence: RWM

- Nonlinear continuous model: We use Proposition 1

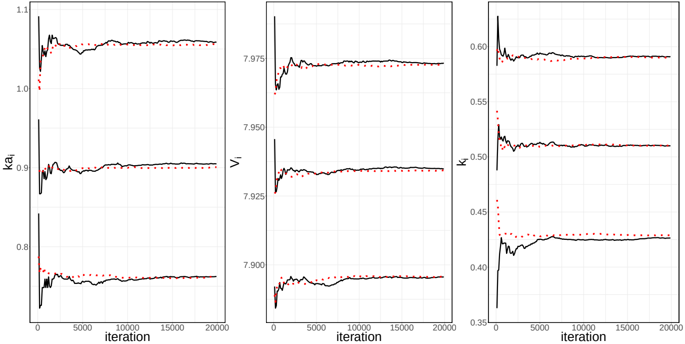

Figure 3: Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $p(\psi_i|y_i;\theta)$ for a single individual. The reference MH algorithm is in black and solid and the new version is in red and dotted.

MCMC convergence: MALA and NUTS

- MALA and NUTS

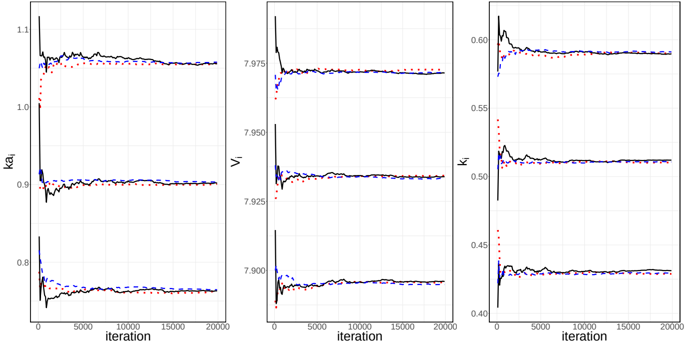

Figure 4: Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathrm{p}(\psi_i|y_i;\theta)$ for a single individual. The new version is in red, the MALA is in black and the NUTS is in blue.

ML Estimation

- With our new proposal (red) versus reference RWM (blue)

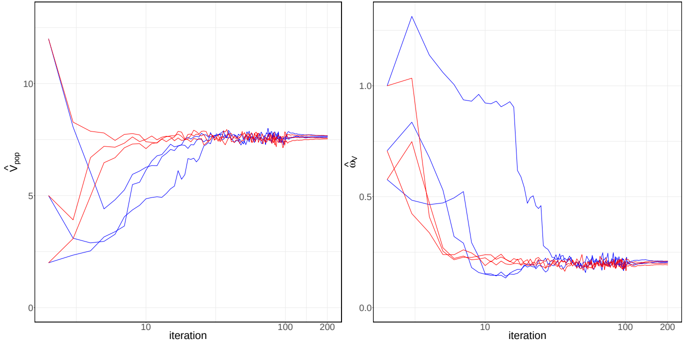

Figure 5: Estimation of the population PK parameters for the warfarin data: convergence of the sequences of estimates $(\hat{V}_{\mathrm{pop},k}, 1 \leq k \leq 200)$ and $(\hat{\omega}_{V,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue) and the new proposal during the first 10 iterations (red).

Monte Carlo Study

- For a given sequence of estimates, we can then define, at each iteration k and for each component ℓ of the parameter, the mean square distance over the replicates

$$E_k(\ell) = \frac{1}{M} \sum_{m=1}^{M} \left(\theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell) \right)^2 . \qquad (14)$$

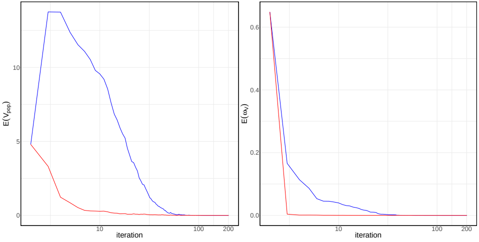

Figure 6: Mean square distances for V_{pop} and ω_V obtained with SAEM on $M = 100$ synthetic datasets using the reference MH algorithm (blue) and the new proposal during the first 10 iterations (red).

Numerical Experiment: Time-To-Event

- Weibull model for time-to-event data [5, 12]. The hazard function of this model for individual i is:

$$h(t, \psi_i) = \frac{\beta_i}{\lambda_i} \left(\frac{t}{\lambda_i} \right)^{\beta_i - 1} . \tag{15}$$

- The vector of individual parameters is $\psi_i = (\lambda_i, \beta_i)$ assumed to be independent and lognormally distributed:

$$\begin{aligned} \log(\lambda_i) &\sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2), \\ \log(\beta_i) &\sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2). \end{aligned} \tag{16}$$

  Then, the vector of population parameters is $\theta = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda^2, \omega_\beta^2)$.
- Individual parameters for $N = 100$ individuals were generated using model (16) with $\lambda_{\text{pop}} = 10$, $\omega_\lambda = 0.3$, $\beta_{\text{pop}} = 3$ and $\omega_\beta = 0.3$. Then, repeated events were generated for each individual using the Weibull model (15) and assuming a right censoring time $\tau_c = 20$.

MCMC convergence: RWM

- Noncontinuous model: We use Proposition 2

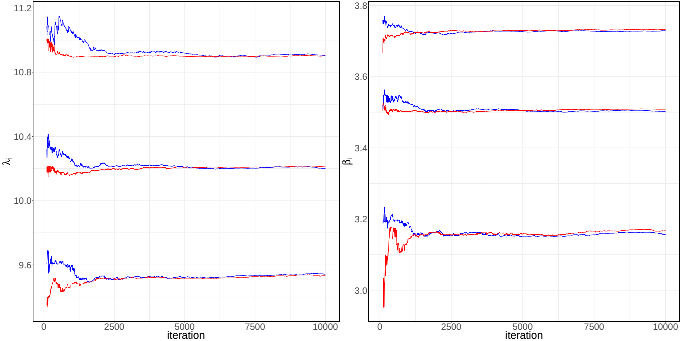

Figure 7: Convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathrm{p}(\psi_i|y_i;\theta)$ for a single individual. The reference MH algorithm is in blue and the new version is in red.

MCMC convergence: MALA

- MALA

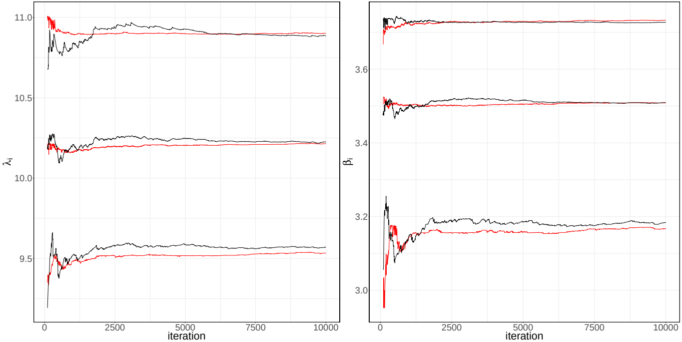

Figure 8: Convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathrm{p}(\psi_i|y_i;\theta)$ for a single individual. The new version is in red, the MALA is in black.

MCMC convergence: NUTS

- NUTS

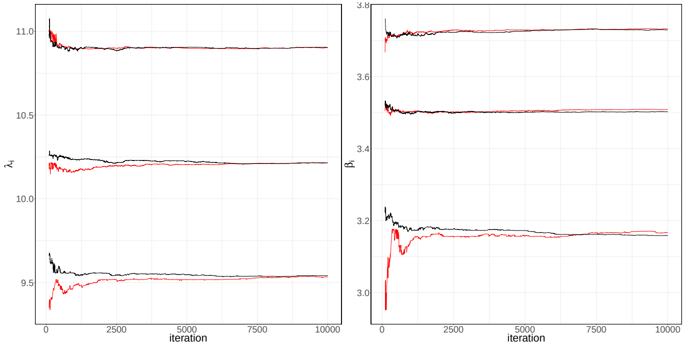

Figure 9: Convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathrm{p}(\psi_i|y_i;\theta)$ for a single individual. The new version is in red, the NUTS is in black.

ML Estimation

- With our new proposal (red) versus reference RWM (blue)

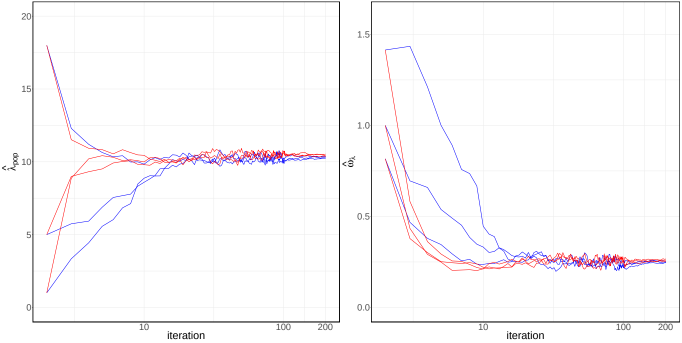

Figure 10: Convergence of the sequences of estimates $(\hat{\lambda}_{\mathrm{pop},k}, 1 \leq k \leq 200)$ and $(\hat{\omega}_{\lambda,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue) and the new proposal during the first 5 iterations (red).

Monte Carlo Study

- Plot of the mean square distance over the replicates

$$E_k(\ell) = \frac{1}{M} \sum_{m=1}^{M} \left(\theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell) \right)^2. \tag{17}$$

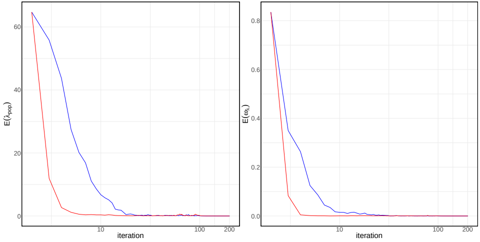

Figure 11: Convergence of the sequences of mean square distances for $\lambda_{\rm pop}$ and ω_λ obtained with SAEM on $M = 100$ synthetic datasets using the reference MH algorithm (blue) and the new proposal during the first 5 iterations (red).

Conclusion

- Automatic and easy to implement proposal for MCMC sampler.
- Leverage the latent structure (random effects here).
- Computing the MAP for each individual is costly: coupling with minibatch strategies (Subsampling MCMC) can be efficient.
- Would be interesting to try on hign dimensional problems.

Thank you!

References

i

Yves F. Atchadé and Jeffrey S. Rosenthal.
On adaptive Markov chain Monte Carlo algorithms.
Bernoulli, 11(5):815–828, 10 2005.

Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Goodrich Ben, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell.
Stan: A probabilistic programming language.
Journal of Statistical Software, 76(1), 2017.

References

ii

Matthew D Hoffman and Andrew Gelman.
The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.
Journal of Machine Learning Research, 15(1):1593–1623, 2014.

E. Kuhn and M. Lavielle.
Coupling a stochastic approximation version of EM with an MCMC procedure.
ESAIM: Probability and Statistics, 8:115–131, 2004.

References

iii

Marc Lavielle.
Mixed effects models for the population approach: models, tasks, methods and tools.
CRC press, 2014.

Yi-An Ma, Tianqi Chen, and Emily Fox.
A complete recipe for stochastic gradient mcmc.
In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2917–2925. Curran Associates, Inc., 2015.

References

iv

Tristan Marshall and Gareth Roberts.
An adaptive approach to langevin mcmc.
Statistics and Computing, 22(5):1041–1057, Sep 2012.

Robert A. O'Reilly and Paul M. Aggeler.
Studies on coumarin anticoagulant drugs initiation of warfarin therapy without a loading dose.
Circulation, 38(1):169–177, 1968.

References

v

Gareth O. Roberts and Richard L. Tweedie.
Exponential convergence of langevin distributions and their discrete approximations.
Bernoulli, 2(4):341–363, 12 1996.

Stan Development Team.
RStan: the R interface to Stan, 2018.
R package version 2.17.3.

References

vi

O. Stramer and R. L. Tweedie.
Langevin-type models i: Diffusions with given stationary distributions and their discretizations*.
Methodology And Computing In Applied Probability, 1(3):283–306, Oct 1999.

Z. Zhang.
Parametric regression model for survival data: Weibull regression model as an example.
Ann Transl Med., 24, 2016.

Appendix

Linear continuous model

- Let $y_i = (y_{i,1}, \ldots, y_{i,n_i})'$ and $\varepsilon_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i})'$.
- Assume first a linear relationship between the observations y_i and the vector of individual parameters ψ_i:

$$y_i = A_i \psi_i + \varepsilon_i, \tag{18}$$

where A_i is the design matrix for individual i and where ψ_i is normally distributed around some value m_i

$$\psi_i \sim \mathcal{N}(m_i\,,\,\Omega).$$

- Then, the conditional distribution of ψ_i is a normal distribution:

$$\psi_i | y_i \sim \mathcal{N}(\mu_i\,,\,\Gamma_i),$$

where

$$\begin{aligned} \Gamma_i &= \left(\frac{A_i' A_i}{\sigma^2} + \Omega^{-1} \right)^{-1}, \\ \mu_i &= \Gamma_i \left(\frac{A_i' y_i}{\sigma^2} + \Omega^{-1} m_i \right). \end{aligned} \tag{19}$$

Proof of Proposition 1

- Defining $z_i = y_i - f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i)\hat{\psi}_i$, the linearization yields

$$z_i = \nabla f_i(\hat{\psi}_i)\psi_i + \varepsilon_i, \tag{20}$$

- We use (19) to get an expression of the conditional variance of ψ_i under the linearized model:

$$\mathrm{Var}_{\mathrm{lin}}(\psi_i|y_i) = \left(\frac{\nabla f_i(\hat{\psi}_i)\nabla f_i(\hat{\psi}_i)'}{\sigma^2} + \Omega^{-1}\right)^{-1}. \tag{21}$$

- The MAP is defined as

$$\hat{\psi}_i = \arg\min_{\psi_i}\left(\frac{1}{\sigma^2}\|y_i - f_i(\psi_i)\|^2 + (\psi_i - m_i)'\Omega^{-1}(\psi_i - m_i)\right),$$

- Thus, $\hat{\psi}_i$ satisfies:

$$-\frac{\nabla f_i(\hat{\psi}_i)'}{\sigma^2}\left(y_i - f_i(\hat{\psi}_i)\right) + \Omega^{-1}(\hat{\psi}_i - m_i) = 0.$$

Proof of Proposition 1

- Let $\Gamma_i = \mathrm{Var}_{\mathrm{lin}}(\psi_i|y_i)$. Using (19), we can now compute the conditional mean of ψ_i under the linearized model:

$$\begin{aligned}
\mathbb{E}_{\mathrm{lin}}(\psi_i|y_i) &= \Gamma_i \frac{\nabla f_i(\hat{\psi}_i)'}{\sigma^2}\left(y_i - f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i)\hat{\psi}_i + \Omega^{-1} m_i\right) \\
&= \Gamma_i \left(\Omega^{-1}(\hat{\psi}_i - m_i) + \frac{\nabla f_i(\hat{\psi}_i)' \nabla f_i(\hat{\psi}_i)}{\sigma^2}\hat{\psi}_i + \Omega^{-1} m_i\right) \\
&= \Gamma_i \Gamma_i^{-1} \hat{\psi}_i \\
&= \hat{\psi}_i.
\end{aligned} \tag{22}$$

Laplace Approximation

- Laplace approximation consists in approximating an integral of the form
$$I := \int e^{v(x)} \mathrm{d}x, \tag{23}$$
where v is at least three times differentiable.
- The following second order Taylor expansion of the function v around a point x_0
$$v(x) \approx v(x_0) + \nabla v(x_0)(x - x_0) + \frac{1}{2}(x - x0)\nabla^2 v(x_0)(x - x0), \tag{24}$$
provides an approximation of the integral I (consider a multivariate Gaussian probability distribution function which integral sums to 1):
$$I \approx e^{v(x_0)} \sqrt{\frac{(2\pi)^p}{|-\nabla^2 v(x_0)|}} \exp\left\{-\frac{1}{2}\nabla v(x_0)'\nabla^2 v(x_0)^{-1}\nabla v(x_0)\right\}. \tag{25}$$

Proof of Proposition 2

- In our context, we can write the marginal pdf $\mathrm{p}(y_i)$ that we aim to approximate as

$$\mathrm{p}(y_i) = \int \mathrm{p}(y_i, \psi_i)\mathrm{d}\psi_i = \int e^{\log(\mathrm{p}(y_i,\psi_i))}\mathrm{d}\psi_i. \tag{26}$$

- Then, let

$$v(\psi_i) = \log(\mathrm{p}(y_i, \psi_i)) = \log(\mathrm{p}(y_i|\psi_i)) + \log(\mathrm{p}(\psi_i)), \tag{27}$$

  and we do the Taylor expansion around the MAP $\hat{\psi}_i$ that verifies by definition $\nabla \log \mathrm{p}(y_i, \hat{\psi}_i) = 0$:

$$-2\log(\mathrm{p}(y_i)) \approx -p\log 2\pi - 2\log \mathrm{p}(y_i, \hat{\psi}_i) + \log\left(\left|-\nabla^2 \log \mathrm{p}(y_i, \hat{\psi}_i)\right|\right).$$

Proof of Proposition 2

- We thus obtain the following approximation of the logarithm of the conditional pdf of ψ_i evaluated at $\hat{\psi}_i$:

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

which is precisely the log-pdf of a multivariate Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance $-\nabla^2 \log \mathrm{p}(y_i, \hat{\psi}_i)^{-1}$, evaluated at $\hat{\psi}_i$, and where

$$\begin{aligned} \nabla^2 \log \mathrm{p}(y_i, \hat{\psi}_i) &= \nabla^2 \log(\mathrm{p}(y_i|\hat{\psi}_i)) + \log(\mathrm{p}(\hat{\psi}_i)) \\ &= \nabla^2 \log(\mathrm{p}(y_i|\psi_i)) + \Omega^{-1}. \end{aligned} \tag{28}$$