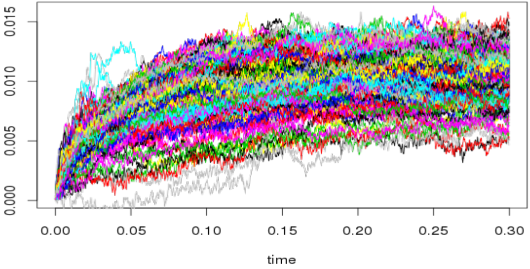
Bayesian and Frequentist algorithms to estimate parameters of Stochastic differential equations

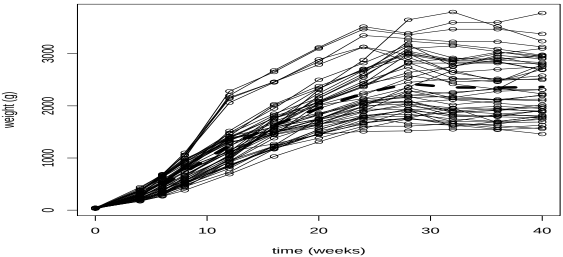
Adeline Leclercq Samson
Joint work with Susanne Ditlevsen (Copenhagen Univ), Sophie Donnet (AgroParisTech), Simone Hermann (Dortmund Univ)

Laboratoire Jean Kuntzmann, Université Grenoble Alpes

Repeated temporal data



Dynamics of neuron voltage



Growth curves

Modeling temporal data/time series

- (Y_{t_k}) at discrete times $(t_k)_{k=1,\dots,n}$
- Noisy observations of a (hidden) temporal process

$$Y_{t_k} = X_{t_k} + \varepsilon_k$$

- Discrete time process (X_{t_k})
    - Regular observation times $t_k = k\Delta$, $X_{t_k} := X_k$
    - Auto-regressive model

$$X_{k+1}|X_k \sim p(\cdot|X_k, \phi)$$

- Continuous time process (X_t)
    - Irregular observation times possible
    - Stochastic differential equation with Brownian motion (B_t)

$$dX_t = a(X_t, \phi)dt + b(X_t, \gamma)dB_t, \quad X_0 = x_0$$

Modeling repeated temporal data/time series

- (Y_{it_k}) at discrete times $(t_k)_{k=1,\dots,n}$ for subject $i = 1, \dots, N$

$$Y_{it_k} = X_{it_k} + \varepsilon_{ik}$$

- Random effects modeling between-subject variability
  - Discrete time process

$$\begin{aligned} X_{ik+1} | X_{ik} &\sim p(\cdot | X_k, \phi_i) \\ \phi_i &\sim p(\varphi, \theta) \end{aligned}$$

  - Continuous time process

$$\begin{aligned} dX_{it} &= a(X_{it}, \phi_i)dt + b(X_{it}, \gamma)dB_{it}, \quad X_{i0} = x_0 \\ \phi_i &\sim p(\varphi, \theta) \end{aligned}$$

Gaussian random effects

$$
\begin{aligned}
y_{ik} &= X_{it_{ik}} + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim_{iid} \mathcal{N}(0, \sigma^2) \\
dX_{it} &= a(X_{it}, \phi_i)dt + b(X_{it}, \gamma)dB_{it}, \quad X_{i0} = x_0 \\
\phi_i &\sim_{iid} \mathcal{N}(\mu, \Omega)
\end{aligned}
$$

Outlines of the talk

- Parameters to be estimated: $\theta = (\mu, \Omega, \gamma, \sigma)$
- Likelihood not explicit $\Rightarrow$ Use of MCMC/PMCMC methods
- Comparison of Bayesian and frequentist approaches

A degenerate Hidden Markov Model

Notations

- $X_i = X_{i,0:n} = (X_{t_{i0}}, \dots, X_{t_{in}})$: hidden diffusion of subject i
- $y_i = y_{i,0:n} = (y_{t_{i0}}, \dots, y_{t_{in}})$: observations of subject i

Why the model is a degenerate HMM

- Hidden coordinates $((X_{it})_{t\geq 0}, i = 1, \dots, N, \phi_{1:N})$
- **Law of observations**
  - conditionally on $((X_{it})_{t\geq 0}, i = 1, \dots, N, \phi_{1:N})$, the (y_{ik}) are independent
- **Law of the hidden process**
  - $(X_{it})_{t\geq 0}$ is a continuous Markov process
  - $((X_{it})_{t\geq 0}, \phi_i)$ has a degenerate dynamics (hypoelliptic system)

$$
\begin{aligned}
dX_{it} &= a(X_{it}, \phi_i)dt + b(X_{it}, \gamma)dB_{it}, \quad X_{i0} = x_0 \\
d\phi_{it} &= 0, \quad \phi_{i0} \sim \mathcal{N}(\mu, \Omega)
\end{aligned}
$$

  - Discrete version $(X_{i,0:n}, \phi_{i,0:n})$ with $\phi_{i,k} = \phi_i$ for all $k = 1, \dots, n$ has a degenerate kernel Q

Alternative

See the model as N independent HMM each with a random parameter ϕ_i

For subject i, the HMM is defined as

- Hidden coordinates $((X_{it})_{t\geq 0})$
- **Law of observations**
  - conditionally on $((X_{it})_{t\geq 0})$, the (y_{ik}) are independent
- **Law of the hidden process**
  - $(X_{it})_{t\geq 0}$ is a continuous Markov process with transition density

$$p(X_{it}|X_{is}, s \leq t; \theta) = \int p(X_{it}|X_{is}, s \leq t; \phi_i) p(\phi_i; \theta) d\phi_i$$

Law of the hidden path $p(X_i;\theta)$

- **Explicit** only for linear drift and known volatility (Girsanov formula)
- **Alternative**: discretization of the SDE with step size Δ (Euler-Maruyama)

$$X_{ik+1} = X_{ik} + \Delta a(X_{ik},\phi_i) + \sqrt{\Delta} b(X_{ik},\gamma)\eta_i, \quad \eta_i \sim_{iid} \mathcal{N}(0,1)$$

- **Approximated transition density**

$$p(X_i;\theta) = \int \prod_{k=0}^{n-1} p(X_{ik+1}|X_{ik},\phi_i;\theta)p(\phi_i;\theta)d\phi_i$$

- **Likelihood**: in any case, integrand explicit but no closed-form of the likelihood

$$p(y_i;\theta) = \int\int \prod_{k=0}^{n} p(y_{ik}|X_{t_{ik}};\theta) \prod_{k=0}^{n-1} p(X_{t_{ik}}|X_{t_{ik-1}},\phi_i;\theta)p(\phi_i;\theta)dX_i d\phi_i$$

Estimation methods

- Exact for Orsntein-Ulhenbeck [Ditlevsen and De Gaetano, 2005]
- Linearization and extended Kalman filter [Tornoe et al, 2005; Overgaard et al 2005, Mortensen et al 2007, Klim et al 2009; Leander et al 2014, 2015]
- Gaussian quadrature [Picchini et al 2010]
- Laplace approximation [Picchini et al 2011]
- **SAEM-MCMC algorithm** [Donnet and Samson, 2008]
- SAEM-Kalman filter [Delattre and Lavielle 2013]
- **SAEM-PMCMC algorithm** [Donnet and Samson, 2014]
- Exact for linear drift [Delattre et al 2013, 15, 16]

- **Bayesian approach with MCMC** [Donnet et al, 2010; Hermann et al 2015]
- **Bayesian approach with PMCMC** [Hermann and Samson, 2016]

Focus on computational methods based on MCMC

Question

Do the MCMC and PMCMC algorithms behave the same in the frequentist and Bayesian approaches ?

Bayesian settings

Hierarchical model with prior distribution

$$\begin{aligned}
y_{ij} &= X_{it_{ij}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim_{iid} \mathcal{N}(0, \sigma^2) \\
dX_{it} &= a(X_{it}, \phi_i)dt + b(X_{it}, \gamma)dB_{it}, \quad X_{i0} = x_0 \\
\phi_i &\sim_{iid} \mathcal{N}(\mu, \Omega) \\
\theta &\sim p(\theta)
\end{aligned}$$

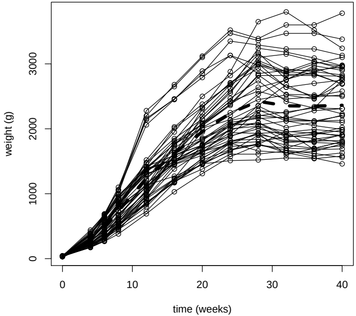
Estimation of $p(\theta|y)$ and of $p(X|y, \theta)$

MCMC: Gibbs algorithm

- **Initialization:** $\theta^{(0)}$, $\phi^{(0)}$, $X^{(0)}$
- **iteration** k**:**

 1. $X^{(k)} \sim p(X|\phi^{(k-1)}, \theta^{(k-1)}, y)$
 2. $\phi^{(k)} \sim p(\phi|\theta^{(k-1)}, X^{(k)}, y)$
 3. $(\mu^{(k)}, \Omega^{(k)}) \sim p((\mu, \Omega)|\phi^{(k)}, X^{(k)}, y)$
 4. $\sigma^{-2(k)} \sim p(\sigma^{-2}|\phi^{(k)}, X^{(k)}, y)$
 5. $\gamma^{-2(k)} \sim p(\gamma^{-2}|\phi^{(k)}, X^{(k)}, y)$

Example with growth curves



Stochastic Gompertz model

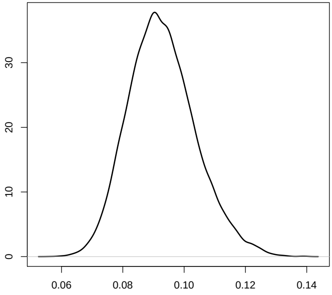
$$
\begin{aligned}
dX_t &= BCe^{-Ct}X_t dt + \gamma X_t dB_t, \\
X_0 &= Ae^{-B}
\end{aligned}
$$

Explicit solution

$$
\begin{aligned}
\log X_{t+\Delta} | (\log X_s)_{s \le t} &\sim \mathcal{N}\left(\log X_t - Be^{-Ct}(e^{-C\Delta} - 1) - \frac{1}{2}\gamma^2\Delta, \gamma^2\Delta\right), \\
\log X_0 &= \log(A) - B
\end{aligned}
$$

Prior and posterior distributions

- Z: Gaussian posterior
- μ: Gaussian prior and posterior
- Ω^{-1}: Wishart prior, inverse Wishart posterior
- σ^{-2}: Gamma prior and posterior
- γ^2: non explicit posterior $\rightarrow$ Metropolis-Hastings



γ^2 posterior distibution

MCMC algorithm: Metropolis-Hastings Within-Gibbs

- 8000 iterations of MCMC
- 5000 iterations of burn-in

Example with neuronal data

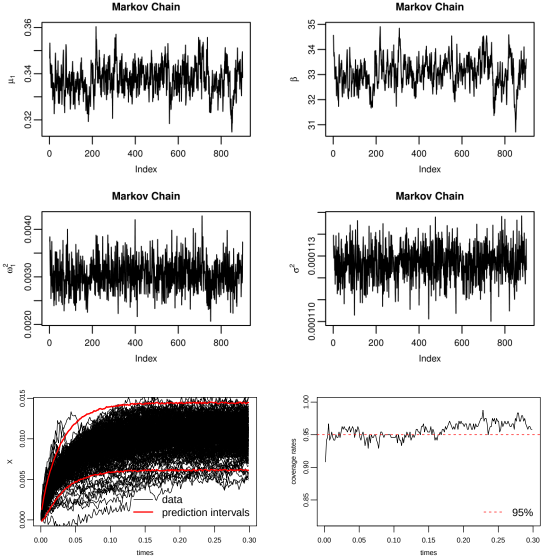
Ornstein-Ulhenbeck process

$$
\begin{aligned}
dX_t &= (\phi_1 - \phi_2 X_t)dt + \gamma dB_t, \\
X_0 &= 0
\end{aligned}
$$

Explicit solution

Standard MCMC algorithm

- 10000 iterations of MCMC
- 9000 iterations of burn-in



Frequentist settings

No prior, maximization of the likelihood

Incomplete data model

- Observed data (y)
- Complete data (y, X, ϕ)

EM Algorithm [Dempster, Laird, Rubin, 1977; Wu, 1983], at iteration m

- *Expectation Step*: calculation of Q_{m+1}

$$Q_{m+1}(\theta) = \mathbb{E}\left[\log p(y, X, \phi; \theta) \,|\, y, \widehat{\theta}_m \right]$$

- *Maximization Step*: update of $\widehat{\theta}_m$

$$\widehat{\theta}_{m+1} = \arg\max_{\theta} Q_{m+1}(\theta)$$

But E step not explicit

Stochastic EM Algorithm

[Celeux Diebolt 1985, Wei Tanner 1990; Delyon, Lavielle and Moulines, 1999]

- *E step*
  - *S step* : simulation of (X_m^ℓ, ϕ_m^ℓ) under distribution $p(X, \phi | y; \widehat{\theta}_m)$ for $\ell = 1, \ldots, L$
  - *E step* : stochastic approximation of Q_{m+1}

$$Q_{m+1}(\theta) = \frac{1}{L} \sum_{\ell=1}^{L} \log p(y, X_m^\ell, \phi_m^\ell; \theta)$$

- *M step* : $\widehat{\theta}_{m+1} = \arg\max_\theta Q_{m+1}(\theta)$

Simulation step with MCMC [Kuhn, Lavielle, 2004]

Convergence results

EM assumptions

1. Incomplete data model in exponential family
2. Regularity of complete likelihood

MCMC assumptions

1. Irreductible and aperiodic Metropolis-Hasting
2. $q(x_j|x_{0:n-1}, y_j, \phi; \theta) > 0$ if $p(x_{0:n}|y_j, \phi; \theta) > 0$

Theorem [Kuhn, Lavielle, 2004]

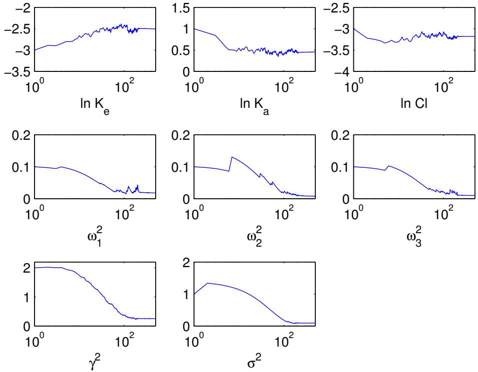
Sequence $(\widehat{\theta}_m)$ converges towards a (local) maximum of likelihood $p(y; \theta)$.

Example with pharmacokinetic data

$$dX_t = \left(\frac{Dose \cdot K_a K_e}{Cl} e^{-K_a t} - K_e X_t \right) dt + \gamma dB_t,$$
$$X_0 = 0$$

Standard MCMC algorithm with Euler-Maruyama approximation

- 700 iterations of EM
- 5 iterations of MCMC at each EM iteration



Problem of the standard MCMC algorithm

- Simulation of the hidden coordinate (X, ϕ)
  - High dimension of X
  - Bad mixing of the chain between X and ϕ

- Apply for Bayesian and frequentist settings

$\Rightarrow$ Natural tool is Particle MCMC [Andrieu et al 2010]
  - Pseudo-marginal MCMC
  - Estimate the likelihood
  - Replace the acceptance ratio with a ratio of estimators

"Ideal" MCMC algorithm

Target distribution: $p(X_{i,0:n}, \phi_i | y_{i,0:n}; \theta)$

- Simulation of $\phi^c \sim q(\cdot; \phi(\ell))$
- Simulation of $X^c \sim p(X_{0:n} | y_{0:n}, \phi^c; \theta)$
- Theoretical acceptance probability

$$\begin{aligned}
\rho &= \min\left\{ \frac{p(X^c, \phi^c | y_{0:n}; \theta)}{p(X(\ell), \phi(\ell) | y_{0:n}; \theta)} \; \frac{q(\phi(\ell); \phi^c)}{q(\phi^c; \phi(\ell))} \, \frac{p(X_{0:n}(\ell) | y_{0:n}, \phi(\ell); \theta)}{p(X^c_{0:n} | y_{0:n}, \phi^c; \theta)}; 1 \right\} \\
&= \min\left\{ \frac{p(\phi(\ell) | y_{0:n}; \theta)}{p(\phi^c | y_{0:n}; \theta)} \; \frac{q(\phi(\ell); \phi^c)}{q(\phi^c; \phi(\ell))}; 1 \right\} \\
&= \min\left\{ \frac{p(y_{0:n} | \phi(\ell); \theta) p(\phi(\ell); \theta)}{p(y_{0:n} | \phi^c; \theta) p(\phi^c; \theta)} \; \frac{q(\phi(\ell); \phi^c)}{q(\phi^c; \phi(\ell))}; 1 \right\}
\end{aligned}$$

- Update of the Markov chain

$$(X(\ell+1), \phi(\ell+1)) = \begin{cases} (X^c, \phi^c) & \text{with proba } \rho \\ (X(\ell), \phi(\ell)) & \text{with proba } 1-\rho \end{cases}$$

Particle filter/Sequential Monte Carlo (SMC)

[Del Moral et al, 2001; Doucet et al, 2001; Chopin, 2004; ...]

Approximate $p(X_{0:n}|y_{0:n},\phi;\theta)$ by K particles $X_{0:n}^{(k)}$ with weights $w\left(X_{0:n}^{(k)}\right)$

$$p\left(X_{0:n}|y_{0:n},\phi;\theta\right) \approx \sum_{k=1}^{K} w\left(X_{0:n}^{(k)}\right) \mathbf{1}_{X_{0:n}^{(k)}}$$

At time $j = 1, \dots, n$, $\forall\, k = 1, \dots, K$:

1. simulation of $X_j^{(k)} \sim q(\cdot|y_j, X_{j-1}^{(k)}, \phi;\theta)$
2. calculation of weights

$$w\left(X_{0:j}^{(k)}\right) = \frac{p(y_{0:j}, X_{0:j}^{(k)}|\phi;\theta)}{p(y_{0:j-1}, X_{0:j-1}^{(k)}|\phi;\theta)\, q(X_j^{(k)}|y_j, X_{j-1}^{(k)}, \phi;\theta)}$$

Conditional likelihood

$$\widehat{p}(y_{0:n}|\phi;\theta) = \widehat{p}(y_0|\phi;\theta) \prod_{j=1}^{n} \widehat{p}(y_j|y_{0:n-1}, \phi;\theta).$$

Particle Marginal Metropolis Hastings (PMMH) algorithm

Approximation of distribution $p(X_{0:n}, \phi_i | y_{0:n}; \theta)$

- *Iteration* $\ell \geq 1$

1. Simulation of $\phi_i^c \sim q(\cdot | \phi_i(\ell - 1))$
2. Simulation of $X_{0:n}^c$ by SMC targeting $p(\cdot | y_{0:n}, \phi_i^c; \theta)$
3. Estimation of $\widehat{p}(y_{0:n} | \phi^c; \theta)$ based on SMC weights with K particles
4. Update

$$(X_{0:n}(\ell), \phi_i(\ell)) = \begin{cases} (X_{0:n}^c, \phi_i^c) & \text{with proba} \quad \widehat{\rho} \\ (X_{0:n}(\ell-1), \phi_i(\ell-1)) & \text{otherwise} \end{cases}$$

where $\widehat{\rho} = \min \left\{ \frac{\widehat{p}(y_{0:n} | \phi(\ell); \theta) p(\phi(\ell); \theta)}{\widehat{p}(y_{0:n} | \phi^c; \theta) p(\phi^c; \theta)} \ \frac{q(\phi(\ell); \phi^c)}{q(\phi^c; \phi(\ell))}; 1 \right\}$

Theorem [Andrieu et al, 2010]

Markov chain produced by PMMH has $p(X, \phi | y; \theta)$ as stationary distribution for any number of particles K

Limits of PMMH

In Stochastic Differential Equations with random parameters

- Hidden stochastic processes: random parameters ϕ and SDE X
- Hyperparameters θ

- PMMH updates simultaneously ϕ and X
  - Low acceptance ratio
  - Poor mixing of the chain

- Alternative: Particle Gibbs Sampler
  - Require conditional SMC [Andrieu et al, 2010]

Particle Gibbs sampler

- *Initialization* : set randomly $\phi^{(0)}$, generate $X^{(0)}$ by a run of SMC algorithm targeting $p(X|, \phi^{(0)}; \gamma, \sigma)$ and store its ancestral lineage $B^{(0)}$.

- *Iteration* $\ell \geq 1$
  1. Sample $\phi^{(\ell)} \sim p(\phi|X^{(\ell-1)}, y)$
  2. Run a conditional SMC algorithm, consistent with $\phi(\ell)$, and the fixed particle $X(\ell-1), B(\ell-1)$, sample $X(\ell)$ from this conditional SMC and denote $B(\ell)$ its ancestral lineage.

Theorem [Andrieu et al, 2010]

Markov chain produced by PGibbs has $p(X, \phi|y; \theta)$ as stationary distribution for any number of particles K

Example with simulated data

Wiener process with drift

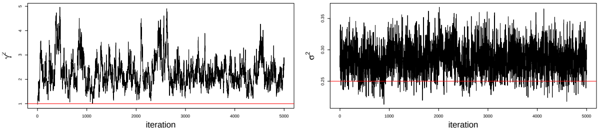
$$
\begin{aligned}
dX_t &= \phi_2 dt + \gamma dB_t, \\
X_0 &= \phi_1
\end{aligned}
$$

PMCMC

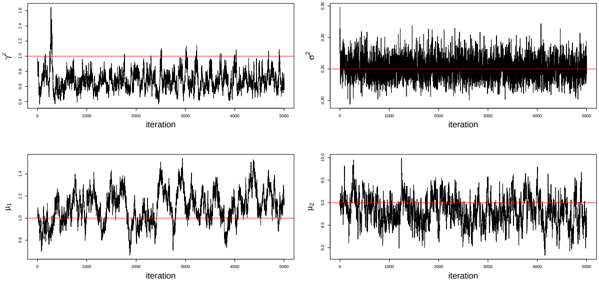
- 5000 iterations
- 1000 iterations of Burn-in
- 100 particles

Comparison of PMMH and PGibbs

PMMH


γ^2
σ^2
μ_1
μ_2
iteration
iteration
iteration
iteration

PGibbs


γ^2
σ^2
μ_1
μ_2
iteration
iteration
iteration
iteration

Coverage rates

	μ_1	μ_2	γ^2	σ^2
Particle Gibbs	0.94	0.95	0.92	0.94
PMMH	0.55	1.00	0.07	0.74

Frequentist settings

Stochastic EM Algorithm coupled with PMCMC

[Donnet, Samson, 2014]

- *E step*
  - *S step* : simulation of (X_m^ℓ, ϕ_m^ℓ) under distribution $p(X, \phi|y; \widehat{\theta}_m)$ with PMCMC using K particles
  - *E step* : stochastic approximation of Q_{m+1}

$$Q_{m+1}(\theta) = \frac{1}{L}\sum_{\ell=1}^{L} \log p(y, X_m^\ell, \phi_m^\ell; \theta)$$

- *M step* : $\widehat{\theta}_{m+1} = \arg\max_{\theta} Q_{m+1}(\theta)$

Theorem [Donnet Samson 2014]

Sequence $(\widehat{\theta}_m)$ converges towards a (local) maximum of likelihood $p(y; \theta)$, for any number of particles with $L = 1$ and stochastic approximation (SAEM)

Stochastic EM and Particle filter

Theoretical results on the minimal number of particles at each iteration

- *E Step*
  - *S Step*:simulation of (X_m, ϕ_m) under distribution $p(X, \phi | y; \widehat{\theta}_m)$ with particle filter using $K(m)$ particles
  - *E step* : stochastic approximation of Q_{m+1}

$$Q_{m+1}(\theta) = \frac{1}{L} \sum_{\ell=1}^{L} \log p(y, X_m^\ell, \phi_m^\ell; \theta)$$

- *M Step*: $\widehat{\theta}_{m+1} = \arg\max_{\theta} Q_{m+1}(\theta)$

Theorem [Ditlevsen, Samson, 2014]

Assumptions: Number of particles $K(m) = \log(m^{1+\delta})$, $L = 1$

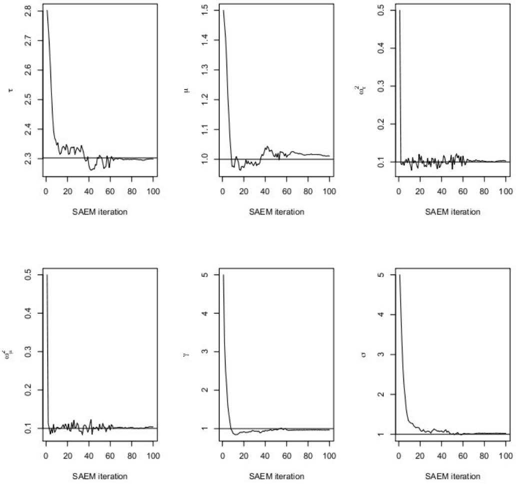
$$\widehat{\theta}_m \xrightarrow[m \to \infty]{a.s.} \textit{(local) max of likelihood}$$

Tool: convergence of Robbins-Monroe scheme and inequality deviation for the particle filter

Example with simulated data

Ornstein-Ulhenbeck process

$$\begin{aligned} dX_t &= -\left(\frac{X_t}{\tau} - \kappa\right) dt + \gamma dB_t, \\ X_0 &= 0 \end{aligned}$$



SAEM-PMMH

- 100 SAEM iterations
- 50 particles
- 10 PMCMC iterations

PMMH tuning

Parameters	$\log(\tau)$	μ	ω_τ	ω_μ	γ	σ
True value	0.600	1.000	0.100	0.100	0.050	0.050
50 particles, 150 PMCMC iterations						
Mean	0.600	1.000	0.096	0.094	0.050	0.050
SD	0.027	0.025	0.026	0.020	0.004	0.002
100 particles, 10 PMCMC iterations						
Mean	0.600	1.000	0.098	0.091	0.051	0.050
SD	0.029	0.028	0.023	0.021	0.004	0.002
50 particles, 10 PMCMC iterations						
Mean	0.600	1.001	0.098	0.094	0.051	0.050
SD	0.027	0.027	0.023	0.019	0.005	0.003
25 particles, 10 PMCMC iterations						
Mean	0.600	1.000	0.097	0.094	0.051	0.050
SD	0.029	0.028	0.023	0.020	0.005	0.003

- Low influence of particle numbers K

Example with simulated data

Stochastic Gompertz process

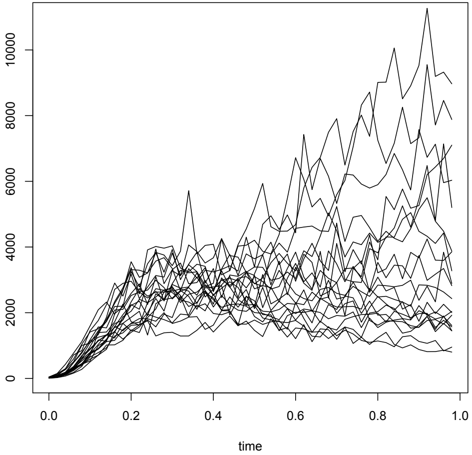
$$
\begin{aligned}
dX_t &= BCe^{-Ct}X_t dt + \gamma X_t dB_t \\
X_0 &= Ae^{-B}
\end{aligned}
$$

SAEM-MCMC

- 100 SAEM iterations
- 100 MCMC iterations

SAEM-PGibbs

- 100 SAEM iterations
- 50 particles
- 100 PGibbs iterations



Results

Parameters		$\log A$	$\log B$	$\log C$	ω_A	ω_B	ω_C	γ	σ
True value		8.006	1.609	2.639	0.100	0.100	0.100	0.500	0.100
SAEM-PGibbs	Mean	8.092	1.623	2.624	0.136	0.097	0.119	0.522	0.100
	SD	0.105	0.024	0.031	0.142	0.013	0.073	0.149	0.004
SAEM-MCMC	Mean	8.522	1.704	2.525	0.073	0.090	0.077	0.746	0.098
	SD	0.252	0.045	0.059	0.028	0.012	0.025	0.118	0.004

- Improvement with SAEM-PGibbs
- Especially γ

Conclusion

- **Time series**
  - Prefer PMCMC for temporal series
  - Particle Gibbs is really better
- **Stochastic Differential Equation with random effects**
  - When transition density unknown: Euler approximation and control of the posterior distribution
  - When irregular observations times: Sampling latent paths [Jenssen et al. 2014]
  - When partially observed SDE