More of the SAME?
Sequential and Pseudomarginal Monte Carlo for
Point Estimation in Latent Variable Models

Adam M. Johansen
Collaborators: Manuel Davy, Arnaud Doucet and Axel Finke

a.m.johnsen@warwick.ac.uk

Imperial College — 6th February, 2014
Outline

- Background:
  - Marginal MLEs
  - SAME: An MCMC Scheme

- Sequential Monte Carlo
  - The SMC Method
  - A Population-Based SAME Method
  - Examples

- Pseudomarginal Methods
  - The Pseudomarginal Method
  - More of the SAME: multiple extensions of the space
  - Example
  - Even more of the SAME: complex extensions of the space
  - Examples
Background

- Marginal MLEs
- SAME: An MCMC Scheme
Consider a model with:
- parameters, \( \theta \),
- latent variables, \( x \), and
- observed data, \( y \).

Aim to maximise marginal likelihood

\[
p(y|\theta) = \int p(x, y|\theta) dx
\]

or posterior

\[
p(\theta|y) \propto \int p(x, y|\theta)p(\theta) dx.
\]

Traditional approach is Expectation-Maximisation (EM)
- Requires objective function in closed form.
- Susceptible to trapping in local optima.
Optimization and probability $\leadsto$ simulated annealing.

A distribution of the form

$$\pi(\theta|y) \propto p(\theta)p(y|\theta)^\gamma$$

will become concentrated, as $\gamma \to \infty$ on the maximisers of $p(y|\theta)$ under weak conditions.

Why not target $\pi(\theta|y)$ using MCMC?
Adapted from (Hwang, 1980; Theorem 2.1).

Assume:

- $p(\theta)$ and $p(y|\theta)$ are $\alpha$-Lipschitz continuous in $\theta$.
- $\log(p(\theta)) \in C^3(\mathbb{R}^n)$ and $\log p(y|\theta) \in C^3(\mathbb{R}^n)$.
- $\Theta_{ML}$ is a non-empty, countable set which is nowhere dense;
- $p(\theta) \leq M < \infty$; $p(\theta) > 0 \forall \theta \in \Theta_{ML}$
- $p(y|\theta) \leq M' < \infty$
- For some $k < \sup p(y|\theta)$, $\{\theta : p(y|\theta) \geq k\}$ is compact.

Then:

$$\lim_{\gamma \to \infty} \pi_\gamma(dt) \propto \sum_{\theta_{ml} \in \Theta_{ML}} \alpha(\theta_{ml}) \delta_{\theta_{ml}}(dt),$$  

(1)

$$\alpha(\theta_{ml}) = \det \left[ - \frac{\partial^2 \log p(y|\theta)}{\partial \theta_m \partial \theta_n} \bigg|_{\theta=\theta_{ml}} \right]^{-1/2}$$  

(2)
Data Augmentation: Synthetic distributions of the form:

\[
\tilde{\pi}_\gamma(\theta, x_{1:\gamma}|y) \propto p(\theta) \prod_{i=1}^{\gamma} p(x_i, y|\theta)
\]

admit the marginals

\[
\tilde{\pi}_\gamma(\theta|y) \propto p(\theta)p(y|\theta)^\gamma.
\]

SAME Algorithm (Doucet, Godsill and Robert, 2002):

- \(t = 0\): Initialise \((\theta_0, X_{0,1})\) arbitrarily.
- For \(t = 1, \ldots, T\):
  - If \(\gamma(t) > \gamma(t-1)\): Set \((X_{t-1,\gamma(t-1)+1}, \ldots, X_{t-1,\gamma(t)})\) arbitrarily.
  - Sample \((\theta_t, X_{t,1}, \ldots, X_{t,\gamma(t)})\) \(\sim K_{\gamma(t)}(\theta_{t-1,1}, X_{t-1,1}, \ldots, X_{t-1,\gamma(t)}, \cdot)\).

  Where \(K_\gamma\) is \(\tilde{\pi}_\gamma\)-invariant.

NB An inhomogeneous Markov chain.
Sequential Monte Carlo

- The SMC Method
- A Population-Based SAME Method
- Examples
- Let $X_1, \ldots$ denote the position of an object which follows Markovian dynamics.
- Let $Y_1, \ldots$ denote a collection of observations:

\[
Y_i | \{X_i = x_i\} \sim g(\cdot | x_i).
\]

- We wish to estimate, as observations arrive, $p(x_{1:t}|y_{1:t})$.
- A recursion obtained from Bayes rule exists but is intractable in most cases.
More Generally

- Really tracking a sequence of distributions, $p_t$...
- on increasing state spaces.
- Other problems with the same structure exist.
- Any problem of sequentially approximating a sequence of such distributions, $p_t$, can be addressed in the same way.
Sequential Importance Resampling

At time $t, t \geq 2$. (Given $\{X_{1:t-1}^{(i)}\}_{i=1}^{N}$ approximating $p_{t-1}(x_{1:t-1})$).

**Sampling Step**

For $i = 1 : N$:

sample $X_t^{(i)} \sim q_t \left( \cdot \mid X_{1:t-1}^{(i)} \right)$.

**Resampling Step**

For $i = 1 : N$:

compute $w_t \left( X_{1:t}^{(i)} \right) = \frac{p_t \left( X_{1:t}^{(i)} \right)}{p_{t-1} \left( X_{1:t-1}^{(i)} \right) q_t \left( X_t^{(i)} \mid X_{1:t-1}^{(i)} \right)}$

and $W_t^{(i)} = \frac{w_t \left( X_{1:t}^{(i)} \right)}{\sum_{j=1}^{N} w_t \left( X_{1:t}^{(j)} \right)}$

For $i = 1 : N$:

sample $A_t^{(i)} \sim \sum_{j=1}^{N} W_t^{(j)} \delta_j$

retain $\{X_{1:t}^{(A_t^{(i)})}\}_{i=1}^{N}$
SMC Samplers (Del Moral et al., 2006)

Can be used to sample from *any* sequence of distributions:

- Given a sequence of *target* distributions, $\eta_n$, on $E_n \ldots$,
- construct a synthetic sequence $\tilde{\eta}_n$ on spaces $\bigotimes_{p=1}^{n} E_p$ by introducing Markov kernels, $L_p$ from $E_{p+1}$ to $E_p$:

$$
\tilde{\eta}_n(x_{1:n}) = \eta_n(x_n) \prod_{p=1}^{n-1} L_p(x_{p+1}, x_p)
$$

- These distributions
  - have the target distributions as final time marginals,
  - have the correct structure to employ SMC techniques.
Given a sample \( \{X_{1:n-1}^{(i)}\}_{i=1}^{N} \) targeting \( \tilde{\eta}_{n-1} \),

sample \( X_n^{(i)} \sim K_n(X_{n-1}^{(i)}, \cdot) \),

calculate

\[
W_n(X_{1:n}^{(i)}) = \frac{\eta_n(X_n^{(i)})L_{n-1}(X_n^{(i)}, X_{n-1}^{(i)})}{\eta_{n-1}(X_{n-1}^{(i)})K_n(X_{n-1}^{(i)}, X_n^{(i)})}.
\]

Resample, yielding: \( \{X_{1:n}^{(i)}\}_{i=1}^{N} \) targeting \( \tilde{\eta}_n \).

Hints that we’d like to use

\[
L_{n-1}(x_n, x_{n-1}) = \frac{\eta_{n-1}(x_{n-1})K_n(x_{n-1}, x_n)}{\int_{x_{n-1}'} \eta_{n-1}(x_{n-1}')K_n(x_{n-1}', x_n)dx_{n-1}}.
\]
A model with:
- parameters, $\theta$,
- latent variables, $x$, and
- observed data, $y$.

Aim to maximise Marginal likelihood

$$p(y|\theta) = \int p(x, y|\theta)dx$$

or posterior

$$p(\theta|y) \propto \int p(x, y|\theta)p(\theta)dx$$

Using

$$\bar{\pi}_\gamma(\theta, x_{1:\gamma}|y) \propto p(\theta)\prod_{i=1}^{\gamma} p(x_i, y|\theta)$$
Maximum Likelihood via SMC

- Use a sequence of distributions \( \eta_n = \bar{\pi}_{\gamma_n} \) for some \( \{\gamma_n\} \).
- The MCMC approach (Doucet et al., 2002).
  - Requires slow “annealing”.
  - Separation between distributions is large.
  - Mixes poorly as \( \gamma \) increases.
- Using SMC has some substantial advantages:
  - Introducing bridging distributions, for \( \gamma = [\gamma] + \langle \gamma \rangle \), of:

\[
\bar{\pi}_{\gamma}(\theta, x_{1:[\gamma]+1}|y) \propto p(\theta)p(x_{[\gamma]+1}, y|\theta)^{\langle \gamma \rangle} \prod_{i=1}^{[\gamma]} p(x_i, y|\theta)
\]

is straightforward.
- Population of samples improves robustness.
- It is less dependent upon mixing of \( K_\gamma \).
A generic SMC sampler can be written down directly...

An easy case:
- Sample from $p(x_t | y, \theta_{t-1})$ and $p(\theta_t | x_t, y)$.
- Weight according to $p(y | \theta_{t-1})^{\gamma_t - \gamma_{t-1}}$.

The general case:
- Sample existing variables from a $\pi_t$-invariant kernel:
  
  $$(\theta_t, X_{t, 1:\gamma_{t-1}}) \sim K_t((\theta_{t-1}, X_{t-1}), \cdot).$$

- Sample new variables from an arbitrary proposal:
  
  $$X_{t, [\gamma_{t-1}+1: \gamma_t]} \sim q(\cdot | \theta_t).$$

- Use combination of time-reversal and optimal auxiliary kernel.
- Weight expression does not involve the marginal likelihood.
initialisation: \( t = 1 \):

sample \( \{(\theta_1^{(i)}, X_1^{(i)}) \sim \nu\}_{i=1}^{N} \)

calculate \( W_1^{(i)} \propto \frac{\pi \gamma_1(\theta_1^{(i)}, X_1^{(i)})}{\nu(\theta_1^{(i)}, X_1^{(i)})} \left( \sum_{i=1}^{N} W_1^{(i)} = 1 \right) \)

for \( t = 2 \) to \( T \) do

resample

\[
\{ \left( \theta_t^{(i)}, X_{t,1:|\gamma_{t-1}|} \right) \sim K_{t-1} \left( \theta_{t-1}^{(i)}, X_{t-1}^{(i)}; \cdot \right) \}
\]

sample \(\{X_t^{(i)} \sim q(\cdot|\theta_t^{(i)}) \}_{j=|\gamma_{t-1}|+1}^{\lceil \gamma_t \rceil}\) if \(|\gamma_{t-1}| < |\gamma_t|\)
\(X_t^{(i)} \sim q_{\langle \gamma_t \rangle}(\cdot|\theta_t^{(i)})\) if \(|\gamma_{t-1}| < |\gamma_t| \neq \gamma_t\)

calculate

\( W_t^{(i)} \propto p(y, X_t, [\gamma_{t-1}]|\theta)^{1 \wedge \gamma_t - |\gamma_{t-1}|} \prod_{j=|\gamma_{t-1}|+1}^{\lceil \gamma_t \rceil} \frac{p(y, X_t, j|\theta_t)}{q(X_t, j|\theta_t)} \left( \frac{p(y, X_t, [\gamma_t] | \theta_t)_{\langle \gamma_t \rangle}}{q_{\langle \gamma_t \rangle}(X_t, [\gamma_t] | \theta_t)} \right)^I \)

with \( I = \mathbb{I}([\gamma_t] > |\gamma_t| \geq |\gamma_{t-1}|) \).
Student $t$-distribution of unknown location parameter $\theta$ with $\nu = 0.05$.

Four observations are available, $y = (-20, 1, 2, 3)$.

Log likelihood is:

$$\log p(y|\theta) = -0.525 \sum_{i=1}^{4} \log \left( 0.05 + (y_i - \theta)^2 \right).$$

Global maximum is at 1.997.

Local maxima at $\{-19.993, 1.086, 2.906\}$.

Complete log likelihood ($X_i \sim Ga$):

$$\log p(y, z|\theta) = - \sum_{i=1}^{4} \left[ 0.475 \log x_i + 0.025 x_i + 0.5 x_i (y_i - \theta)^2 \right].$$
Toy Example: Marginal Likelihood

Toy Example: Log Marginal Likelihood

log marginal likelihood vs $\theta$

The graph shows the log marginal likelihood as a function of $\theta$. The likelihood peaks sharply near $\theta = 0$, with secondary peaks further away.
Toy Example: SMC Method using Gibbs Kernels
• Likelihood $p(y|x, \omega, \mu, \sigma) = \mathcal{N}(y|\mu_x, \sigma_x^2)$.

• Marginal likelihood $p(y|\omega, \mu, \sigma) = \sum_{j=1}^{3} \omega_j \mathcal{N}(y|\mu_j, \sigma_j^2)$.

• Diffuse conjugate priors were employed:

  $\omega \sim \text{Di}(\delta)$

  $\sigma_i^2 \sim \text{IG} \left( \frac{\lambda_i + 3}{2}, \frac{\beta_i}{2} \right)$

  $\mu_i|\sigma_i^2 \sim \mathcal{N} \left( \alpha_i, \frac{\sigma_i^2}{\lambda_i} \right)$,

• All full conditional distributions of interest are available.

• Marginal posterior can be calculated.
3 Component GMM (Roeders Galaxy Data Set)
Pseudomarginal Monte Carlo

- The Pseudomarginal Method
- More of the SAME: multiple extensions of the space
- Example
- Even more of the SAME: complex extensions of the space
- Examples
The Pseudomarginal Method

- Marginal MH-Acceptance Probability:

\[ 1 \wedge \frac{\pi(\theta')Q(\theta', \theta)}{\pi(\theta)Q(\theta, \theta')} \]

- But \( \pi(\theta) \) isn’t tractable: how about using:

\[ 1 \wedge \frac{\hat{\pi}(\theta')Q(\theta', \theta)}{\hat{\pi}(\theta)Q(\theta, \theta')} \]

where

\[ \hat{\pi}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{\pi(\theta, X_i)}{q(X_i)} \quad X_i \overset{iid}{\sim} q \]

- Suggests two algorithms (Beaumont, 2003):
  - Monte Carlo within Metropolis
  - Grouped Independence Metropolis Hastings
Pseudomarginal Methods: GIMH is “Exact”

- Extended Target (Andrieu & Roberts, 2009):

\[
\tilde{\pi}(\theta, x_1, \ldots, x_m) = \sum_{j=1}^{m} \frac{1}{m} \pi(\theta, x_j) \prod_{k \neq j} q(x_k) \\
= \frac{1}{m} \sum_{j=1}^{m} \frac{\pi(\theta, x_j)}{q(x_j)} \cdot \prod_{k=1}^{m} q(x_k) = \hat{\pi}(\theta) \prod_{k=1}^{m} q(x_k)
\]

- The acceptance probability becomes:

\[
1 \land \frac{\tilde{\pi}(\theta', x'_1, \ldots, x'_m)Q(\theta', \theta) \prod_{j=1}^{m} q(x'_j)}{\tilde{\pi}(\theta, x_1, \ldots, x_m)Q(\theta, \theta') \prod_{j=1}^{m} q(x_j)} = 1 \land \frac{\hat{\pi}(\theta')Q(\theta', \theta)}{\hat{\pi}(\theta)Q(\theta, \theta')}
\]

- NB MCWM is not exact... but perhaps we don’t care.
We’d like to target $\pi_\gamma(\theta|y) \propto p(\theta)p(y|\theta)^\gamma$.

Why not use the pseudomarginal approach, considering instead:

$$\tilde{\pi}_\gamma(\pi, x^1_{1:m}, \ldots, x^\gamma_{1:m}) = p(\theta) \prod_{i=1}^{\gamma} \sum_{j=1}^{m} \frac{1}{m} \frac{p(x^i_j, y|\theta)}{q(x^i_j|\theta)} \prod_{k=1}^{m} q(x^i_k|\theta)$$

Expect behaviour like simulated annealing for large $m$. 
Summary of 200 Runs

$N = 5$

$N = 50$

$N = 100$

$\theta$

GIMH (prior)  GIMH  MCWM (prior)  MCWM

GIMH (prior)  GIMH  MCWM (prior)  MCWM

GIMH (prior)  GIMH  MCWM (prior)  MCWM

SAME  SA
What about complicated latent variable structures?

- Actually, pseudomarginal algorithms are more flexible.
- We’re especially interested in particle MCMC implementations (Andrieu et al., 2010):
  - Particle Marginal Metropolis-Hastings (PMMH)
  - MCWM variant of PMMH
  - Particle Gibbs (with ancestor sampling)
- State-space models are the real motivation for this methodology.
- Many other complex models could be addressed using this technique.
Model:

\[ X_t = AX_{t-1} + BU_t \]
\[ Y_t = X_t + DV_t \]

Data 50 observations simulated using:

\[ A = 0.9, B = 1, \text{ and } D = 1 \]

Algorithms

- The PMMH/MCWM algorithms use \( N = 250 \) particles;
- The PG algorithm (with ancestor sampling) uses \( N = 50 \) particles but attempts 100 static parameter updates per iteration.
- Inverse temperature increases linearly from 0.1 to 10.
- Final 1000 iterations \( \gamma_t = 10 \).
- Compare with exact marginal simulated annealing algorithm.
Summary of 100 Runs

A

B

C

D

PMMH (N = 250)
MCWM (N = 250)
Particle Gibbs (N = 50)
Simulated Annealing

PMMH (N = 250)
MCWM (N = 250)
Particle Gibbs (N = 50)
Simulated Annealing

PMMH (N = 250)
MCWM (N = 250)
Particle Gibbs (N = 50)
Simulated Annealing
A Simple Stochastic Volatility Model

Model:

\[ X_i = \alpha + \delta X_{i-1} + \sigma_u u_i \quad \text{with} \quad X_1 \sim \mathcal{N}(\mu_0, \sigma_0^2) \]

\[ Y_i = \exp \left( \frac{X_i}{2} \right) \epsilon_i \]

where \( u_i \) and \( \epsilon_i \) are uncorrelated standard normal random variables, and \( \theta = (\alpha, \delta, \sigma_u) \).

- 200 Observations; simulated with \( \delta = 0.95 \), \( \alpha = -0.363 \) and \( \sigma = 0.26 \).
- Diffuse instrumental prior distributions:
  - \( \delta \sim U(-1, 1) \)
  - \( \alpha \sim \mathcal{N}(0, 1) \)
  - \( \sigma^{-2} \sim \mathcal{Ga}(1, 0.1) \)

are quickly forgotten.

- Inverse temperature increases linearly from 0.1 to 10.
- Final 500 iterations \( \gamma_t = 10 \).
- A more complex multi-factor model is also under investigation.
Monte Carlo isn’t just for calculating posterior expectations.
SMC and Pseudomarginal methods are effective for ML and MAP estimation.
Still work in progress...
Scope for embedding Pseudomarginal target within SMC algorithm...
and adaptation.


