

Particles as auxiliary variables: PMCMC, SMC², and all that jazz

N. CHOPIN¹

CREST-ENSAE

¹based on joint work(s) with S. BARTHELME, C. SCHÄFER, P.E. JACOB,
O. PAPASPILIOPOULOS & S.S. SINGH

Outline

- 1 Introduction
- 2 GIMH
- 3 PMCMC
- 4 SMC²
- 5 Conclusion

the whole SMC vs MCMC debate...

- MCMC is more versatile
- SMC is more specialised, but better at what it does
- PMCMC (and related approaches) \approx MCMC+SMC: best of both worlds?

Main theme of this talk

At first glance, PMCMC and related methods may be understood as: some magic which allows to replace in MCMC an **intractable** quantity by an **unbiased** Monte Carlo estimate, while preserving the validity of the approach.

However, this is only half of the story:

- various variations and extensions of PMCMC cannot be derived from this unbiasedness property only: Particle Gibbs, SMC², etc.
- Worse, unbiasedness is not sufficient: a ‘blind’ replacement may lead to invalid algorithms.

Of course, an ‘invalid’ algorithm may remain valid in a ‘double asymptotics’ sense (number of MCMC iterations, and number of MC samples used at each iteration, both goes to infinity), but this is much less appealing.

Outline

- 1 Introduction
- 2 GIMH**
- 3 PMCMC
- 4 SMC²
- 5 Conclusion

General framework

To fix ideas, consider joint distribution $\pi(\theta, x)$, where x is often much bigger than θ , and

- 1 either we are interested only in $\pi(\theta) = \int \pi(\theta, x) dx$:
- 2 or we would like to construct a sampler that would behave almost as well as a "good" marginal sampler.

In Bayesian Statistics, one typically has:

$$\pi(\theta, x) \propto p(\theta)p(x|\theta)p(y|x, \theta)$$

GIMH

In addition, consider the following **unbiased** estimator of $\pi(\theta)$:

$$\hat{\pi}(\theta) = \frac{1}{N} \sum_{n=1}^N \frac{\pi(\theta, x^n)}{q(x^n)}, \quad x^{1:N} \stackrel{iid}{\sim} q(x)$$


GIMH (Beaumont, 2003) is Metropolis with $\pi(\theta)$ replaced by $\hat{\pi}(\theta)$:

GIMH

From current point θ_m

- 1 Sample $\theta_* \sim T(\theta_m, \theta_*)$
- 2 With probability $1 \wedge r$, take $\theta_{m+1} = \theta_*$, otherwise $\theta_{m+1} = \theta_m$, where

$$r = \frac{\hat{\pi}(\theta_*) T(\theta_*, \theta_m)}{\hat{\pi}(\theta_m) T(\theta_m, \theta_*)}$$

Is GIMH a **non-standard** HM sampler w.r.t. **standard** target $\pi(\theta)$? 

Validity of GIMH

Property 1

The following quantity

$$\bar{\pi}(\theta, x^{1:N}) = \prod_{n=1}^N q(x^n) \hat{\pi}(\theta)$$

is a joint probability density, whose θ -marginal is $\pi(\theta)$.

Proof: Direct consequence of unbiasedness; fix θ then

$$\int \prod_{n=1}^N q(x^n) \hat{\pi}(\theta) dx^{1:N} = \mathbb{E} [\hat{\pi}(\theta)] = \pi(\theta)$$

GIMH as a Metropolis sampler

Property 2

GIMH is a Metropolis sampler with respect to joint distribution $\bar{\pi}(\theta, x^{1:N})$. The proposal is $T(\theta_m, \theta_*) \prod_{n=1}^N q(x_m^n)$.

Proof: current point is $(\theta_m, x_m^{1:N})$, proposed point is $(\theta_*, x_*^{1:N})$ and HM ratio is

$$r = \frac{\prod_{n=1}^N q(x_*^n) \hat{\pi}(\theta_*) T(\theta_*, \theta_m) \prod_{n=1}^N q(x_m^n)}{\prod_{n=1}^N q(x_m^n) \hat{\pi}(\theta_m) T(\theta_m, \theta_*) \prod_{n=1}^N q(x_*^n)}$$

Thus, GIMH is a **standard** Metropolis sampler w.r.t. **non-standard** (extended) target $\bar{\pi}(\theta, x^{1:N})$.

There is more to life than this

Property 3

Extend $\bar{\pi}(\theta, x^{1:N})$ with $k|\theta, x^{1:N} \propto \pi(\theta, x^k)/q(x^k)$, then,

- the marginal dist. of (θ, x^k) is $\pi(\theta, x)$.
- Conditional on (θ, x^k) , $x_n \sim q$ for $n \neq k$, independently.

Proof: let

$$\bar{\pi}(\theta, x^{1:N}, k) = \left\{ \prod_{n=1}^N q(x^n) \right\} \frac{\pi(\theta, x^k)}{q(x^k)} = \left\{ \prod_{n \neq k} q(x^n) \right\} \pi(\theta, x^k)$$

then clearly the sum w.r.t. k gives $\bar{\pi}(\theta, x^{1:N})$, while the above properties hold.

Main lessons

- Unbiasedness (of $\hat{p}(\theta)$) has just been used as an intermediate result.
- Unbiasedness does not provide any intuition on Proposition 3; i.e. (a) how to sample not only θ , but (θ, x) ; and (b) ability to do Gibbs sampling.
- Finally, unbiasedness does not necessary lead to a valid algorithm.

Unbiasness without an auxiliary variable representation

This time, consider instead a target $\pi(\theta)$ (no x), involving an intractable **denominator**; an important application is Bayesian inference on likelihoods with intractable normalising constants:

$$\pi(\theta) \propto p(\theta)p(y|\theta) = p(\theta) \frac{h_\theta(y)}{Z(\theta)}$$

Liang & Lin (2010)'s sampler

From current point θ_m

- 1 Sample $\theta_\star \sim T(\theta_m, \theta_\star)$
- 2 With probability $1 \wedge r$, take $\theta_{m+1} = \theta_\star$, otherwise $\theta_{m+1} = \theta_m$, where

$$r = \left(\frac{\widehat{Z(\theta_m)}}{Z(\theta_\star)} \right) \frac{p(\theta_\star)h_{\theta_\star}(y)T(\theta_\star, \theta_m)}{p(\theta_m)h_{\theta_m}(y)T(\theta_m, \theta_\star)}.$$

Outline

- 1 Introduction
- 2 GIMH
- 3 PMCMC**
- 4 SMC²
- 5 Conclusion

PMCMC: introduction

PMCMC (Andrieu et al., 2010) is akin to GIMH, except a more complex proposal mechanism is used: a PF (particle filter). Thus, the same remarks will apply:

- Unbiasness (of the likelihood estimated provided by the PF) is only an intermediate result for establishing the validity of the whole approach.
- Unbiasness is not enough to give you intuition on the validity of e.g. Particle Gibbs.

State Space Models

A system of equations

- Hidden states (Markov): $p(x_1|\theta) = \mu_\theta(x_1)$ and for $t \geq 1$

$$p(x_{t+1}|x_{1:t}, \theta) = p(x_{t+1}|x_t, \theta) = f_\theta(x_{t+1}|x_t)$$

- Observations:

$$p(y_t|y_{1:t-1}, x_{1:t-1}, \theta) = p(y_t|x_t, \theta) = g_\theta(y_t|x_t)$$

- Parameter: $\theta \in \Theta$, prior $p(\theta)$. We observe $y_{1:T} = (y_1, \dots, y_T)$, T might be large ($\approx 10^4$). x and θ will also be of several dimensions.

There are several interesting models for which f_θ cannot be written in closed form (but it can be simulated).

State Space Models

Some interesting distributions

Bayesian inference focuses on:

$$\text{static: } p(\theta|y_{1:T}) \quad \text{dynamic: } p(\theta|y_{1:t}), t \in 1 : T$$

Filtering/Smoothing (traditionally) focus on ($\forall t \in 1 : T$):

$$p_{\theta}(x_t|y_{1:t}) \quad p_{\theta}(x_t|y_{1:T})$$

More challenging:

$$\text{static: } p(\theta, x_{1:T}|y_{1:T}) \quad \text{dynamic: } p(\theta, x_{1:t}|y_{1:t}), t \in 1 : T$$

Examples

Population growth model

$$\begin{cases} y_t & = x_t + \sigma_w \varepsilon_t \\ \log x_{t+1} & = \log x_t + b_0 + b_1(x_t)^{b_2} + \sigma_\epsilon \eta_t \end{cases}$$

$$\theta = (b_0, b_1, b_2, \sigma_\epsilon, \sigma_w).$$

Examples

Stochastic Volatility (Lévy-driven models)

- Observations (“log returns”):

$$y_t = \mu + \beta v_t + v_t^{1/2} \epsilon_t, t \geq 1$$

- Hidden states (“actual volatility” - integrated process):

$$v_{t+1} = \frac{1}{\lambda} \left(z_t - z_{t+1} + \sum_{j=1}^k e_j \right)$$

Examples

... where the process z_t is the “spot volatility”:

$$z_{t+1} = e^{-\lambda} z_t + \sum_{j=1}^k e^{-\lambda(t+1-c_j)} e_j$$

$$k \sim \text{Poi}(\lambda \xi^2 / \omega^2) \quad c_{1:k} \stackrel{iid}{\sim} U(t, t+1) \quad e_{i:k} \stackrel{iid}{\sim} \text{Exp}(\xi / \omega^2)$$

The parameter is $\theta \in (\mu, \beta, \xi, \omega^2, \lambda)$, and $x_t = (v_t, z_t)'$.

▶ See the results

Why are those models challenging?

... It is effectively impossible to compute the likelihood

$$p(y_{1:T}|\theta) = \left[\int_{\mathcal{X}^T} p(y_{1:T}|x_{1:T}, \theta) p(x_{1:T}|\theta) dx_{1:T} \right]$$

Similarly, all other inferential quantities are impossible to compute.

Problems with MCMC approaches

- Metropolis-Hastings:
 - ① $p(\theta|y_{1:T})$ cannot be evaluated point-wise (marginal MH)
 - ② $p(x_{1:T}, \theta|y_{1:T})$ are high-dimensional and it is hard to design reasonable proposals
- Gibbs sampler (updates states and parameters):
 - ① The hidden states $x_{1:T}$ are typically very correlated and it is hard to update them efficiently in a block
 - ② Parameters and latent variables highly correlated
- Common: they are not designed to recover the whole sequence $p(x_{1:t}, \theta | y_{1:t})$ for $t \in 1 : T$.

Particle filters

Consider the simplified problem of targeting

$$p_{\theta}(x_{t+1}|y_{1:t+1})$$

This sequence of distributions is approximated by a sequence of **weighted particles** which are **properly weighted** using importance sampling, mutated/propagated according to the system dynamics, and resampled to control the variance.

Below we give a pseudo-code version. Any operation involving the superscript n must be understood as performed for $n = 1 : N_x$, where N_x is the total number of particles.

Step 1: At iteration $t = 1$,

- (a) Sample $x_1^n \sim q_{1,\theta}(\cdot)$.
- (b) Compute and normalise weights

$$w_{1,\theta}(x_1^n) = \frac{\mu_\theta(x_1^n) g_\theta(y_1 | x_1^n)}{q_{1,\theta}(x_1^n)}, \quad W_{1,\theta}^n = \frac{w_{1,\theta}(x_1^n)}{\sum_{n=1}^N w_{1,\theta}(x_1^i)}.$$

Step 2: At iteration $t = 2 : T$

- (a) Sample the index $a_{t-1}^n \sim \mathcal{M}(W_{t-1}^{1:N_x})$ of the ancestor
- (b) Sample $x_t^n \sim q_{t,\theta}(\cdot | x_{t-1}^{a_{t-1}^n})$.
- (c) Compute and normalise weights

$$w_{t,\theta}(x_{t-1}^{a_{t-1}^n}, x_t^n) = \frac{f_\theta(x_t^n | x_{t-1}^{a_{t-1}^n}) g_\theta(y_t | x_t^n)}{q_{t,\theta}(x_t^n | x_{t-1}^{a_{t-1}^n})}, \quad W_{t,\theta}^n = \frac{w_{t,\theta}(x_{t-1}^{a_{t-1}^n}, x_t^n)}{\sum_{n=1}^{N_x} w_{t,\theta}(x_{t-1}^{a_{t-1}^i}, x_t^i)}$$

Particle filtering

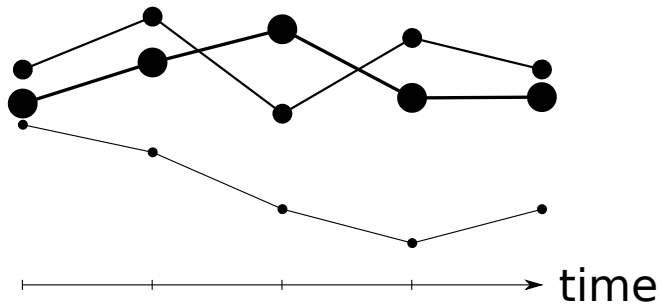


Figure: Three weighted trajectories $x_{1:t}$ at time t .

Particle filtering

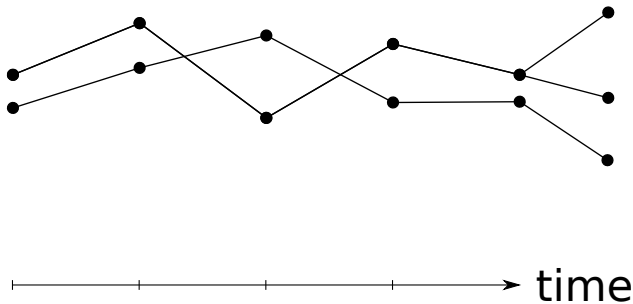


Figure: Three proposed trajectories $x_{1:t+1}$ at time $t + 1$.

Particle filtering

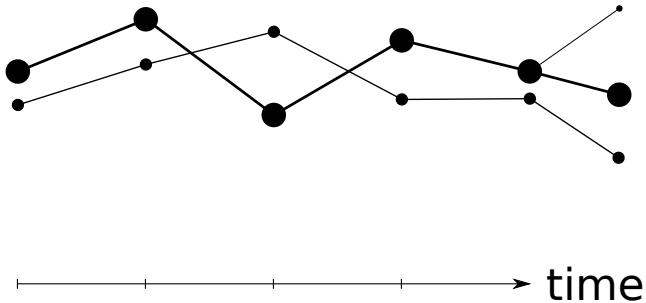


Figure: Three reweighted trajectories $x_{1:t+1}$ at time $t + 1$

Observations

- At each t , $(w_t^n, x_t^n)_{n=1}^{N_x}$ is a particle approximation of $p_\theta(x_t | y_{1:t})$.
- Resampling to avoid degeneracy. If there were no interaction between particles there would be typically polynomial or worse increase in the variance of weights
- Taking $q_\theta = f_\theta$ simplifies weights, but mainly yields a feasible algorithm when f_θ can only be simulated.

Unbiased likelihood estimator

A by-product of PF output is that

$$\hat{Z}_t^N = \prod_{t=1}^T \left(\frac{1}{N_x} \sum_{n=1}^{N_x} w_t^{(i)} \right)$$

is an **unbiased** estimator of the likelihood $Z_t(\theta) = p(y_{1:t}|\theta)$ for all t .

(Not trivial, see e.g Proposition 7.4.1 in Pierre Del Moral's book.)
The variance of this estimator grows typically linealy with T .

PMCMC

Breakthrough paper of Andrieu et al. (2011), based on the unbiasedness of the PF estimate of the likelihood.

Marginal PMCMC

From current point θ_m (and current PF estimate $\hat{p}(y|\theta_m)$):

- 1 Sample $\theta_\star \sim T(\theta_m, d\theta_\star)$
- 2 Run a PF so as to obtain $\hat{p}(y|\theta_\star)$, an unbiased estimate of $p(y|\theta_\star)$.
- 3 With probability $1 \wedge r$, set $\theta_{m+1} = \theta_\star$, otherwise $\theta_{m+1} = \theta_m$ with

$$r = \frac{p(\theta_\star)\hat{p}(y|\theta_\star)T(\theta_\star, \theta_m)}{p(\theta_m)\hat{p}(y|\theta_m)T(\theta_m, \theta_\star)}$$

Validity

Property 1

Let $\psi_{t,\theta}(x_{1:t}^{1:N}, a_{1:t-1}^{1:N})$ be the joint distribution of all the the random variables generated by a PF (for a fixed θ) up to time t , then the following quantity

$$\pi_t(\theta, x_{1:t}^{1:N}, a_{1:t-1}^{1:N}) = \frac{p(\theta)}{p(y_{1:t})} \psi_{t,\theta}(x_{1:t}^{1:N}, a_{1:t-1}^{1:N}) \hat{p}(\theta|y_{1:t})$$

is a pdf, such that the θ -marginal is $p(\theta|y_{1:t})$.

Proof: unbiasedness

$$\begin{aligned} \int \pi_t(\cdot) d(x_{1:t}^{1:N}, a_{1:t-1}^{1:N}) &= \frac{p(\theta)}{p(y_{1:t})} \mathbb{E} [\hat{p}(y_{1:t}|\theta)] \\ &= \frac{p(\theta)p(y_{1:t}|\theta)}{p(y_{1:t})} = p(\theta|y_{1:t}) \end{aligned}$$

More direct proof for $T = 2$ (and $q_\theta = f_\theta$)

$$\psi_{2,\theta}(x_{1:2}^{1:N}, a_1^{1:N}) = \prod_{n=1}^N \mu_\theta(x_1^n) \left\{ \prod_{n=1}^N f_\theta(x_2^n | x_1^{a_1^n}) W_{1,\theta}^{a_1^n} \right\}$$

with $W_{1,\theta}^n = w_{1,\theta}(x_1^n) / \sum_{n=1}^N w_{1,\theta}(x_1^n)$, $w_{1,\theta}(x_1) = g_\theta(y_1 | x_1)$. So

$$\pi_2(\theta, x_{1:2}^{1:N}, a_1^{1:N}) = \frac{p(\theta)}{p(y_{1:t})} \psi_{2,\theta}(\cdot) \left\{ \frac{1}{N} \sum_{n=1}^N w_{1,\theta}(x_1^n) \right\} \left\{ \frac{1}{N} \sum_{n=1}^N w_{2,\theta}(x_2^n) \right\}$$

$$= \frac{p(\theta)}{N^2 p(y_{1:t})} \sum_{n=1}^N g_\theta(y_2 | x_2^n) f_\theta(x_2^n | x_1^{a_1^n}) \frac{g_\theta(x_1^{a_1^n})}{\sum_{n=1}^N w_{1,\theta}(x_1^n)} \left\{ \sum_{n=1}^N w_{1,\theta}(x_1^n) \right\}$$

$$\times \mu_\theta(x_1^{a_1^n}) \left\{ \prod_{i \neq a_1^n} f_{1,\theta}(x_1^i) \right\} \left\{ \prod_{i \neq n} f_\theta(x_2^i | x_1^{a_1^i}) W_{1,\theta}^{a_1^i} \right\}$$

Interpretation

$$\pi_2(\theta, x_{1:2}^{1:N}, a_1^{1:N}) = \frac{1}{N} \times \frac{1}{N} \sum_{n=1}^N p(\theta, x_1^{a_1^n}, x_2^n | y_{1:2}) \prod_{i \neq n} \mu_\theta(x_1^i) \left\{ \prod_{i \neq n} f_\theta(x_2^i | x_1^{a_1^i}) W_1^{a_1^i} \right\}$$

which is a mixture distribution, with probability $1/N$ that path n follows $p(\theta, x_{1:2} | y_{1:2})$, and other paths follows a conditional SMC distribution (and a_1^n is Uniform in $1 : N$).

From this calculation, one easily deduce the unbiasedness property (directly!) but also properties similar to those of the GIMH.

Additional properties (similar to GIMH)

- If we add component $k \in 1 : N$ with conditional distribution $\propto W_2^k$, then the joint pdf $\pi_2(\theta, x_{1:2}^{1:N}, a_1^{1:N}, k)$ is such that (a) $(\theta, x_1^{a_1^k}, x_2^k)$ follows the target distribution $p(\theta, x_{1:2} | y_{1:2})$; and (b) the $N - 1$ remaining trajectories admit some conditional dist. known as the conditional SMC distribution.
- Marginal PMCMC is a Metropolis sampler with invariant distribution $\pi_2(\theta, x_{1:2}^{1:N}, a_1^{1:N})$, and proposal distribution $T(\theta, \theta_*) \psi_{2, \theta_*}(x_{1:2}^{1:N}, a_1^{1:N})$.
- We can do a Gibbs step, by re-generating the $N - 1$ trajectories that differ from trajectory k .

Note on resampling schemes

Calculations are easier when the standard multinomial scheme is considered, but most results carry over to any resampling scheme that is **marginally unbiased** (that is, the marginal probability that $a_t^i = j$ is W_t^j).

The conditional SMC step is slightly more involved when alternative resampling schemes are used, but (up to some simple modifications), it is still doable, and seems to lead to better mixing properties (ongoing work with Sumeet Singh).

PMCMC Summary

- As announced, unbiasedness is not sufficient to motivate theoretically PMCMC, nor to give intuition on how it works.
- In fact, direct calculations obtain unbiasedness as a by-product, rather than an useful result per se.

Outline

- 1 Introduction
- 2 GIMH
- 3 PMCMC
- 4 SMC²**
- 5 Conclusion

Preliminary

So far, we have played with replacing intractable quantities with unbiased estimated within Metropolis samplers. Note however we could do the same within an importance sampler. For instance, the following approach has been used in Chopin and Robert (2007).

To compute the evidence $p(y)$ of some state-space model

- Sample points θ^n from the prior $p(\theta)$.
- For each θ^n , run a PF (for fixed $\theta = \theta^n$) to obtain an estimate $\hat{p}(y|\theta^n)$ of the likelihood.
- Compute

$$\hat{p}(y) = \frac{1}{N} \sum_{n=1}^N \hat{p}(y|\theta^n)$$

Objectives

- 1 to derive sequentially

$$p(\theta, x_{1:t}|y_{1:t}), \quad p(y_{1:t}), \quad \text{for all } t \in \{1, \dots, T\}$$

- 2 to obtain a **black box** algorithm (automatic calibration).

Main tools of our approach

- Particle filter algorithms for state-space models (this will be to estimate the likelihood, for a fixed θ).
- Iterated Batch Importance Sampling for sequential Bayesian inference for parameters (this will be the theoretical algorithm we will try to approximate).

Both are sequential Monte Carlo (SMC) methods

IBIS

SMC method for particle approximation of the sequence $p(\theta \mid y_{1:t})$ for $t = 1 : T$. PF is not going to work here by just pretending that θ is a dynamic process with zero (or small) variance. Recall the path degeneracy problem.

In the next slide we give the pseudo-code of the IBIS algorithm. Operations with superscript m must be understood as operations performed for all $m \in 1 : N_\theta$, where N_θ is the total number of θ -particles.

Sample θ^m from $p(\theta)$ and set $\omega^m \leftarrow 1$. Then, at time $t = 1, \dots, T$

- (a) Compute the incremental weights and their weighted average

$$u_t(\theta^m) = p(y_t | y_{1:t-1}, \theta^m), \quad L_t = \frac{1}{\sum_{m=1}^{N_\theta} \omega^m} \times \sum_{m=1}^{N_\theta} \omega^m u_t(\theta^m),$$

- (b) Update the importance weights,

$$\omega^m \leftarrow \omega^m u_t(\theta^m). \quad (1)$$

- (c) If some degeneracy criterion is fulfilled, sample $\tilde{\theta}^m$ independently from the mixture distribution

$$\frac{1}{\sum_{m=1}^{N_\theta} \omega^m} \sum_{m=1}^{N_\theta} \omega^m K_t(\theta^m, \cdot).$$

Finally, replace the current weighted particle system:

$$(\theta^m, \omega^m) \leftarrow (\tilde{\theta}^m, 1).$$

Observations

- Cost of lack of ergodicity in θ : the occasional MCMC move
- Still, in regular problems resampling happens at diminishing frequency (logarithmically)
- K_t is an MCMC kernel invariant wrt $\pi(\theta | y_{1:t})$. Its parameters can be chosen using information from current population of θ -particles
- L_t is a MC estimator of the **model evidence**
- Infeasible to implement for state-space models: intractable incremental weights, and MCMC kernel

Our algorithm: SMC²

We provide a generic (black box) algorithm for recovering the sequence of parameter posterior distributions, but as well filtering, smoothing and predictive.

We give next a pseudo-code; the code seems to only track the parameter posteriors, but actually it does all other jobs.

Superficially, it looks an approximation of IBIS, but in fact it **does not produce any systematic errors** (unbiased MC).

Sample θ^m from $p(\theta)$ and set $\omega^m \leftarrow 1$. Then, at time $t = 1, \dots, T$,

- (a) For each particle θ^m , perform iteration t of the PF: If $t = 1$, sample independently $x_1^{1:N_x, m}$ from ψ_{1, θ^m} , and compute

$$\hat{p}(y_1 | \theta^m) = \frac{1}{N_x} \sum_{n=1}^{N_x} w_{1, \theta}(x_1^{n, m});$$

If $t > 1$, sample $(x_t^{1:N_x, m}, a_{t-1}^{1:N_x, m})$ from ψ_{t, θ^m} conditional on $(x_{1:t-1}^{1:N_x, m}, a_{1:t-2}^{1:N_x, m})$, and compute

$$\hat{p}(y_t | y_{1:t-1}, \theta^m) = \frac{1}{N_x} \sum_{n=1}^{N_x} w_{t, \theta}(x_{t-1}^{a_{t-1}^{n, m}, m}, x_t^{n, m}).$$

(b) Update the importance weights,

$$\omega^m \leftarrow \omega^m \hat{p}(y_t | y_{1:t-1}, \theta^m)$$

(c) If some degeneracy criterion is fulfilled, sample $(\tilde{\theta}^m, \tilde{x}_{1:t}^{1:N_x, m}, \tilde{a}_{1:t-1}^{1:N_x})$ independently from

$$\frac{1}{\sum_{m=1}^{N_\theta} \omega^m} \sum_{m=1}^{N_\theta} \omega^m K_t \left\{ \left(\theta^m, x_{1:t}^{1:N_x, m}, a_{1:t-1}^{1:N_x, m} \right), \cdot \right\}$$

Finally, replace current weighted particle system:

$$(\theta^m, x_{1:t}^{1:N_x, m}, a_{1:t-1}^{1:N_x, m}, \omega^m) \leftarrow (\tilde{\theta}^m, \tilde{x}_{1:t}^{1:N_x, m}, \tilde{a}_{1:t-1}^{1:N_x, m}, 1)$$

Observations

- It appears as approximation to IBIS. For $N_x = \infty$ it is IBIS.
- However, no approximation is done whatsoever. This algorithm really samples from $p(\theta|y_{1:t})$ and all other distributions of interest. One would expect an increase of MC variance over IBIS.
- The validity of algorithm is essentially based on two results: i) the particles are **weighted** due to unbiasedness of PF estimator of likelihood; ii) the MCMC kernel is appropriately constructed to maintain invariance wrt to an **expanded distribution** which admits those of interest as marginals; it is a **Particle MCMC kernel**.
- The algorithm does not suffer from the path degeneracy problem due to the MCMC updates

The MCMC step

- (a) Sample $\tilde{\theta}$ from proposal kernel, $\tilde{\theta} \sim T(\theta, d\tilde{\theta})$.
- (b) Run a new PF for $\tilde{\theta}$: sample independently $(\tilde{x}_{1:t}^{1:N_x}, \tilde{a}_{1:t-1}^{1:N_x})$ from $\psi_{t,\tilde{\theta}}$, and compute $\hat{Z}_t(\tilde{\theta}, \tilde{x}_{1:t}^{1:N_x}, \tilde{a}_{1:t-1}^{1:N_x})$.
- (c) Accept the move with probability

$$1 \wedge \frac{\rho(\tilde{\theta}) \hat{Z}_t(\tilde{\theta}, \tilde{x}_{1:t}^{1:N_x}, \tilde{a}_{1:t-1}^{1:N_x}) T(\tilde{\theta}, \theta)}{\rho(\theta) \hat{Z}_t(\theta, x_{1:t}^{1:N_x}, a_{1:t-1}^{1:N_x}) T(\theta, \tilde{\theta})}$$

It can be shown that this is a standard Hastings-Metropolis kernel with proposal

$$q_{\theta}(\tilde{\theta}, \tilde{x}_{1:t}^{1:N_x}, \tilde{a}_{1:t}^{1:N_x}) = T(\theta, \tilde{\theta}) \psi_{t,\tilde{\theta}}(\tilde{x}_{1:t}^{1:N_x}, \tilde{a}_{1:t}^{1:N_x})$$

invariant wrt to an extended distribution $\pi_t(\theta, x_{1:t}^{1:N_x}, a_{1:t-1}^{1:N_x})$.

Some advantages of the algorithm

- Immediate estimates of filtering and predictive distributions
- Immediate and sequential estimator of model evidence
- Easy recovery of smoothing distributions
- Principled framework for automatic calibration of N_x
- Population Monte Carlo advantages

Validity

SMC² is simply a SMC sampler with respect to the sequence:

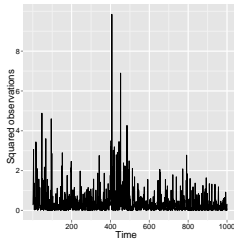
$$\pi_t(\theta, x_{1:t}^{1:N}, a_{1:t-1}^{1:N})$$

- the reweighting step $t \rightarrow t + 1$ (a) extends the dimension, by sampling $x_{t+1}^{1:N}, a_t^{1:N}$; and (b) computes $\pi_{t+1}(\cdot)/\pi_t(\cdot)$.
- The move step is a PMCMC step that leaves π_t invariant.

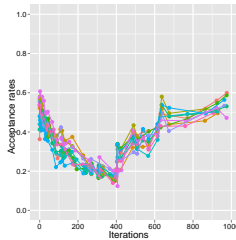
Technical point

As in PMCMC, one may extend π_t by adding index k that picks some trajectory, which, jointly with θ , is sampled from the current posterior $p(\theta, x_{1:t}|y_{1:t})$. However, it is more difficult to define an importance sampling step with respect to the extended space (that includes k), so, we must discard k before progressing to time $t + 1$.

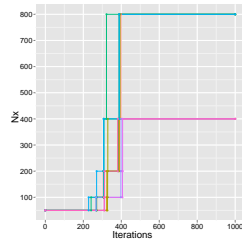
Numerical illustrations: SV



(a)



(b)



(c)

Figure: Squared observations (synthetic data set), acceptance rates, and illustration of the automatic increase of N_x .

▶ See the model

Numerical illustrations: SV

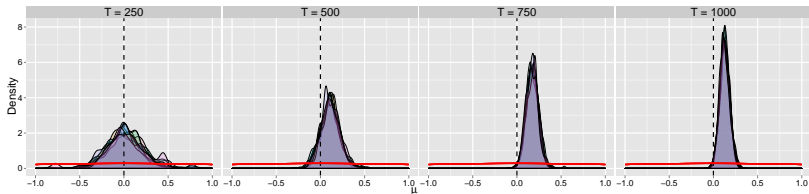


Figure: Concentration of the posterior distribution for parameter μ .

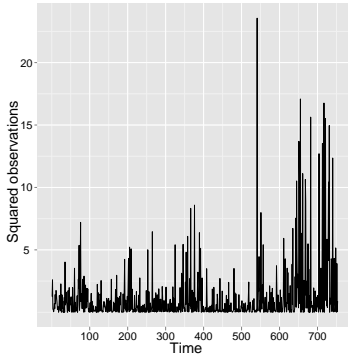
Numerical illustrations: SV

Multifactor model

$$y_t = \mu + \beta v_t + v_t^{1/2} \epsilon_t + \rho_1 \sum_{j=1}^{k_1} e_{1,j} + \rho_2 \sum_{j=1}^{k_2} e_{2,j} - \xi(w\rho_1\lambda_1 + (1-w)\rho_2\lambda_2)$$

where $v_t = v_{1,t} + v_{2,t}$, and $(v_i, z_i)_{n=1,2}$ are following the same dynamics with parameters $(w_i\xi, w_i\omega^2, \lambda_i)$ and $w_1 = w$, $w_2 = 1 - w$.

Numerical illustrations: SV



(a)



(b)

Figure: S&P500 squared observations, and log-evidence comparison between models (relative to the one-factor model).

Numerical illustrations

Athletics records model

$$g(y_{1:2,t}|\mu_t, \xi, \sigma) = \{1 - G(y_{2,t}|\mu_t, \xi, \sigma)\} \prod_{n=1}^2 \frac{g(y_{i,t}|\mu_t, \xi, \sigma)}{1 - G(y_{i,t}|\mu_t, \xi, \sigma)}$$

$$x_t = (\mu_t, \dot{\mu}_t)', \quad x_{t+1} | x_t, \nu \sim \mathcal{N}(Fx_t, Q),$$

with

$$F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } Q = \nu^2 \begin{pmatrix} 1/3 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

$$G(y|\mu, \xi, \sigma) = 1 - \exp \left[- \left\{ 1 - \xi \left(\frac{y - \mu}{\sigma} \right) \right\}_+^{-1/\xi} \right]$$

Numerical illustrations

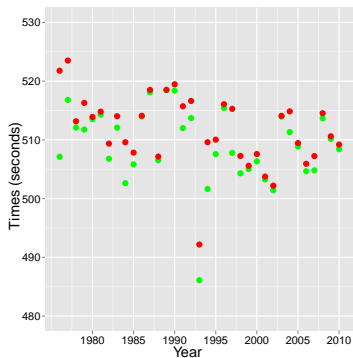


Figure: Best two times of each year, in women's 3000 metres events between 1976 and 2010.

Numerical illustrations: Athletics records

Motivating question

How unlikely is Wang Junxia's record in 1993?

A smoothing problem

We want to estimate the likelihood of Wang Junxia's record in 1993, given that we observe a better time than the previous world record. We want to use all the observations from 1976 to 2010 to answer the question.

Note

We exclude observations from the year 1993.

▶ See the model

Numerical illustrations

Some probabilities of interest

$$\begin{aligned} p_t^y &= \mathbb{P}(y_t \leq y | y_{1976:2010}) \\ &= \int_{\Theta} \int_{\mathcal{X}} G(y | \mu_t, \theta) p(\mu_t | y_{1976:2010}, \theta) p(\theta | y_{1976:2010}) d\mu_t d\theta \end{aligned}$$

The interest lies in $p_{1993}^{486.11}$, $p_{1993}^{502.62}$ and $p_t^{cond} := p_t^{486.11} / p_t^{502.62}$.

Numerical illustrations

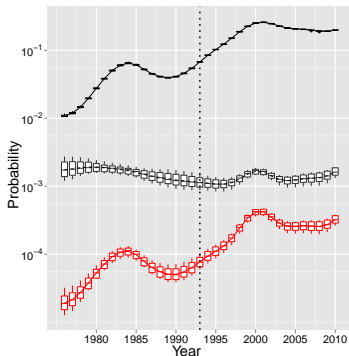


Figure: Estimates of the probability of interest (top) $p_t^{502.62}$, (middle) p_t^{cond} and (bottom) $p_t^{486.11}$, obtained with the SMC² algorithm. The y-axis is in log scale, and the dotted line indicates the year 1993 which motivated the study.

Final Remarks on SMC²

A powerful framework

- A **generic** algorithm for sequential estimation and state inference in state space models: only requirements are to be able (a) to simulate the Markov transition $f_{\theta}(x_t|x_{t-1})$, and (b) to evaluate the likelihood term $g_{\theta}(y_t|x_t)$.
- The article is available on arXiv and our web pages
- A package is available at:

<http://code.google.com/p/py-smc2/>.

Outline

- 1 Introduction
- 2 GIMH
- 3 PMCMC
- 4 SMC²
- 5 Conclusion**

General conclusions

- Auxiliary variables algorithms are not so complicated, when they are understood as **standard** samplers on **extended** spaces.
- offers excellent performance, at little cost (in the user's time dimension); almost magic.
- Many applications not yet fully explored; e.g. variable selection, see C. Schäfer's PhD. thesis.
- Many avenues for future research, e.g. the active particle framework of Anthony Lee (work with Arnaud and Christophe).

Appendix

Why does it work? - Intuition for $t = 1$

At time $t = 1$, the algorithm generates variables θ^m from the prior $p(\theta)$, and for each θ^m , the algorithm generates vectors $x_1^{1:N_x, m}$ of particles, from $\psi_{1, \theta^m}(x_1^{1:N_x})$.

Thus, the sampling space is $\Theta \times \mathcal{X}^{N_x}$, and the actual “particles” of the algorithm are N_θ independent and identically distributed copies of the random variable $(\theta, x_1^{1:N_x})$, with density:

$$p(\theta)\psi_{1,\theta}(x_1^{1:N_x}) = p(\theta) \prod_{n=1}^{N_x} q_{1,\theta}(x_1^n).$$

Then, these particles are assigned importance weights corresponding to the incremental weight function $\hat{Z}_1(\theta, x_1^{1:N_x}) = N_x^{-1} \sum_{n=1}^{N_x} w_{1,\theta}(x_1^n)$.

This means that, at iteration 1, the target distribution of the algorithm should be defined as:

$$\pi_1(\theta, x_1^{1:N_x}) = p(\theta) \psi_{1,\theta}(x_1^{1:N_x}) \times \frac{\hat{Z}_1(\theta, x_1^{1:N_x})}{p(y_1)},$$

where the normalising constant $p(y_1)$ is easily deduced from the property that $\hat{Z}_1(\theta, x_1^{1:N_x})$ is an unbiased estimator of $p(y_1|\theta)$.

Direct substitutions yield

$$\begin{aligned}\pi_1(\theta, x_1^{1:N_x}) &= \frac{p(\theta)}{p(y_1)} \prod_{n=1}^{N_x} q_{1,\theta}(x_1^n) \left\{ \frac{1}{N_x} \sum_{n=1}^{N_x} \frac{\mu_\theta(x_1^n) g_\theta(y_1 | x_1^n)}{q_{1,\theta}(x_1^n)} \right\} \\ &= \frac{1}{N_x} \sum_{n=1}^{N_x} \frac{p(\theta)}{p(y_1)} \mu_\theta(x_1^n) g_\theta(y_1 | x_1^n) \left\{ \prod_{n=1, i \neq n}^{N_x} q_{1,\theta}(x_1^i) \right\}\end{aligned}$$

and noting that, for the triplet (θ, x_1, y_1) of random variables,

$$p(\theta) \mu_\theta(x_1) g_\theta(y_1 | x_1) = p(\theta, x_1, y_1) = p(y_1) p(\theta | y_1) p(x_1 | y_1, \theta)$$

one finally gets that:

$$\pi_1(\theta, x_1^{1:N_x}) = \frac{p(\theta | y_1)}{N_x} \sum_{n=1}^{N_x} p(x_1^n | y_1, \theta) \left\{ \prod_{n=1, i \neq n}^{N_x} q_{1,\theta}(x_1^i) \right\}.$$

By a simple induction, one sees that the target density π_t at iteration $t \geq 2$ should be defined as:

$$\pi_t(\theta, x_{1:t}^{1:N_x}, a_{1:t-1}^{1:N_x}) = p(\theta) \psi_{t,\theta}(x_{1:t}^{1:N_x}, a_{1:t-1}^{1:N_x}) \times \frac{\hat{Z}_t(\theta, x_{1:t}^{1:N_x}, a_{1:t-1}^{1:N_x})}{p(y_{1:t})}$$

and the following Proposition

Proposition

The probability density π_t may be written as:

$$\begin{aligned} \pi_t(\theta, \mathbf{x}_{1:t}^{1:N_x}, \mathbf{a}_{1:t-1}^{1:N_x}) &= p(\theta | y_{1:t}) \\ &\times \frac{1}{N_x} \sum_{n=1}^{N_x} \frac{p(\mathbf{x}_{1:t}^n | \theta, y_{1:t})}{N_x^{t-1}} \left\{ \prod_{\substack{n=1 \\ i \neq \mathbf{h}_t^n(1)}}^{N_x} q_{1,\theta}(x_1^i) \right\} \\ &\times \left\{ \prod_{s=2}^t \prod_{\substack{n=1 \\ i \neq \mathbf{h}_t^n(s)}}^{N_x} W_{s-1,\theta}^{a_{s-1}^i} q_{s,\theta}(x_s^i | x_{s-1}^{a_{s-1}^i}) \right\} \end{aligned}$$