Sequential Monte Carlo with estimated likelihoods

Richard Everitt

University of Reading

March 4th, 2016

Collaborators

- Noisy MCMC: Pierre Alquier (ENSAE ParisTech), Nial Friel and Aidan Boland (UCD).
- Noisy IS and SMC: Adam Johansen (Warwick), Melina Evdemon-Hogan and Ellen Rowing (Reading).
- Recent SMC work: Philip Maybank (Reading), Dennis Prangle (Newcastle).

Papers

- Everitt R. G. (2012). Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks, Journal of Computational and Graphical Statistics, 21(4), 940-960, or arXiv(1203.3725)
- Alquier, P., Friel, N., Everitt, R. G., Boland, A. (2015). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels, Statistics and Computing, or arXiv(1403.5496).
- Everitt, R. G., Johansen, A. M., Rowing, E., Evdemon-Hogan, M. (2016). Bayesian model comparison with un-normalised likelihoods, arXiv(1504.00298).

Bayesian inference

- θ is unknown.
- y is data.

$$\begin{aligned} \pi(\theta|y) &= \frac{p(\theta)f(y|\theta)}{p(y)} \\ &\propto p(\theta)f(y|\theta). \end{aligned}$$

The marginal likelihood

- The marginal likelihood (also known as the evidence) is

$$p(y) = \int_\theta p(\theta) f(y|\theta) d\theta.$$

- Used in Bayesian model comparison

$$p(M|y) = p(M)p(y|M),$$

most commonly seen in the Bayes' factor, for comparing models

$$\frac{p(y|M_1)}{p(y|M_2)}.$$

Importance sampling (IS)

Importance sampling

Returns a weighted sample $\{(\theta^{(p)}, w^{(p)}) \mid 1 \leq p \leq P\}$ from $\pi(\theta|y)$.

- For $p = 1 : P$
  - Simulate $\theta^{(p)} \sim q(.)$
  - Weight $\widetilde{w}^{(p)} = \frac{p(\theta^{(p)})f(y|\theta^{(p)})}{q(\theta^{(p)})}$.

- Then

$$\widehat{\mathbb{E}[\theta]} = \sum_{p=1}^{P} w^{(p)}\theta^{(p)} \qquad \widehat{p}(y) = \frac{1}{P}\sum_{p=1}^{P} \widetilde{w}^{(p)}.$$

Types of intractable likelihood

- A likelihood is intractable when it is difficult to evaluate pointwise at θ.

1. Big data
$$f(y|\theta) = \prod_{i=1}^{N} f_i(y_i|\theta).$$
2. When there are a large number of latent variables x, with
$$f(y|\theta) = \int_x f(y,x|\theta)dx.$$
3. When, for an intractable $Z(\theta)$ (e.g for a *Markov random field*),
$$f(y|\theta) = \frac{1}{Z(\theta)}\gamma(y|\theta).$$
4. Where $f(\cdot|\theta)$ can be sampled, but not evaluated.

Main approach

- For each θ, it is possible to compute an estimate $\widehat{f}(y|\theta)$ of $f(y|\theta)$.
- Includes:
  - approximate Bayesian computation (ABC);
  - synthetic likelihood (SL);
  - psuedo-marginal methods (including particle MCMC);
  - emulators;
  - composite likelihood;
  - many others...

Exact-approximate methods

- Suppose that, for any θ, it is possible to compute an unbiased estimate $\widehat{f}(y|\theta)$ of $f(y|\theta)$. Then...

1. Using the acceptance probability

$$\alpha\left(\theta^{(p)},\theta^{*}\right)=\min\left\{1,\frac{\widehat{f}(y|\theta^{*})p(\theta^{*})q(\theta^{(p)}|\theta^{*})}{\widehat{f}(y|\theta^{(p)})p(\theta^{(p)})q(\theta^{*}|\theta^{(p)})}\right\}$$

yields an MCMC algorithm with target distibution $\pi(\theta|y)$.

2. Using the weight

$$w^{(p)}=\frac{\widehat{f}(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})}$$

yields an importance sampling algorithm with target distribution $\pi(\theta|y)$.

Beaumont (2003), Andrieu and Roberts (2009), Fearnhead et al. (2010).

Why is this true?

- Write down the joint distrubution of *all* of the variables that are being used
$$\widehat{f}(y|\theta,u)p(u|\theta)p(\theta)$$
where u are the random variables used to generate the estimate $\widehat{f}$.
- An algorithm that simulates from $\pi(\theta,u|y)$ has the correct marginal
$$\begin{aligned}
\int_u \pi(\theta,u|y)du &\propto \int_u \widehat{f}(y|\theta,u)p(u|\theta)p(\theta)du \\
&= p(\theta)\int_u \widehat{f}(y|\theta,u)p(u|\theta)du \\
&= p(\theta)f(y|\theta) \\
&\propto \pi(\theta|y).
\end{aligned}$$

Why is this true?

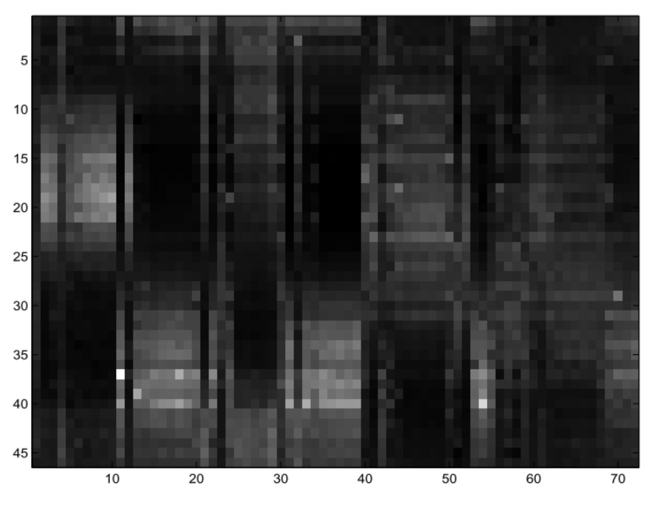
- Using $q\left(\theta^{(p)}\right) p\left(u^{(p)}|\theta^{(p)}\right)$ as a proposal within an importance sampling algorithm yields the desired importance weight.

$$\frac{\widehat{f}(y|\theta^{(p)}, u^{(p)})p(u^{(p)}|\theta^{(p)})p(\theta^{(p)})}{q\left(\theta^{(p)}\right) p\left(u^{(p)}|\theta^{(p)}\right)}$$

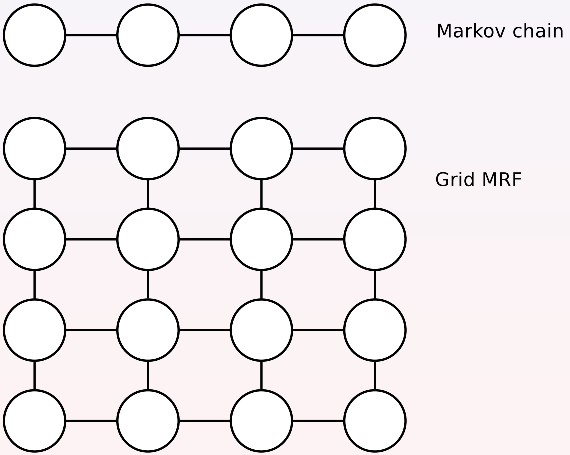
$$= \frac{\widehat{f}(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})}.$$

- A similar extended space representation may be used in MCMC.

Noisy images



Pairwise Markov random fields



Ising models

- Originally used as a model for ferromagnetism in statistical physics.
- Generalisations (including the *Potts model*) are frequently used in analysing spatially structured data, especially images.
- A pairwise factorisation on a grid, where each variable can take on either the value -1 or 1.
- Each potential is:

$$\Phi(x_i, x_j | \theta_x) = \exp(\theta_x x_i x_j), \tag{1}$$

so that the joint distribution is:

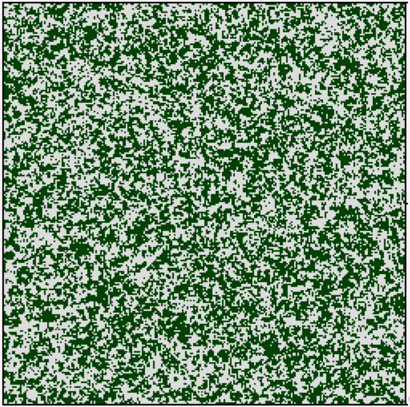
$$f(x|\theta_x) = \frac{1}{Z(\theta_x)} \exp\left(\theta_x \sum_{i,j} (x_{i,j} x_{i,j+1} + x_{i,j} x_{i+1,j})\right). \tag{2}$$

- So a larger parameter results in neighbouring variables being likely to be similar.

Ising models

- Models undergo a phase transition as θ_x increases:

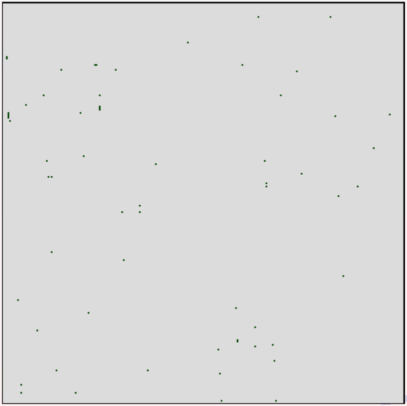
Figure: θ_x just lower than the critical value.



Ising models

- Models undergo a phase transition as θ_x increases:

Figure: θ_x just greater than the critical value.



Type 3: "doubly intractable" distributions

- Coined by Murray et al. (2006).
- Doubly intractable since the acceptance probability in MH

$$\min\left\{1, \frac{\gamma(y|\theta^*)}{\gamma(y|\theta^{(p)})}\frac{p(\theta^*)}{p(\theta^{(p)})}\frac{q(\theta^{(p)}|\theta^*)}{q(\theta^*|\theta^{(p)})}\frac{1}{Z(\theta^*)}\frac{Z(\theta^{(p)})}{1}\right\}$$

requires evaluating the intractable term Z.

- Often take the form

$$f(y|\theta) = \frac{1}{Z(\theta)}\exp\left(\theta^T S(y)\right).$$

Importance sampling for marginal likelihoods

- Importance sampling:

$$
\begin{aligned}
p(y) &= \int_{\theta} \frac{f(y|\theta)p(\theta)}{q(\theta)} q(\theta) d\theta \\
&\approx \frac{1}{P} \sum_{p=1}^{P} \frac{f(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})} \\
&= \frac{1}{P} \sum_{p=1}^{P} \frac{\gamma(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})} \frac{1}{Z(\theta^{(p)})}.
\end{aligned}
$$

- Intractable...

Importance sampling for marginal likelihoods

- Importance sampling:

$$\begin{aligned} p(y) &= \int_\theta \frac{f(y|\theta)p(\theta)}{q(\theta)} q(\theta) d\theta \\ &\approx \frac{1}{P} \sum_{p=1}^{P} \frac{f(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})} \\ &= \frac{1}{P} \sum_{p=1}^{P} \frac{\gamma(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})} \frac{1}{Z(\theta^{(p)})}. \end{aligned}$$

- Intractable...

SAV importance sampling

- Everitt et al. (2016) use

$$\frac{1}{Z(\theta^*)} \approx \frac{q_u(u^*|\theta^*, y)}{\gamma(u^*|\theta^*)}$$

with some distribution q_u and $u^* \sim f(.|\theta^*)$.

- Using $\frac{q_u(u|\theta^*,y)}{\gamma(u|\theta^*)}$ as an IS estimator of $\frac{1}{Z(\theta^*)}$ we obtain

$$w^{(p)} = \frac{\gamma(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})} \frac{q_u(u|\theta^{(p)}, y)}{\gamma(u|\theta^{(p)})}.$$

- Note: we may use multiple importance points, i.e. use

$$\frac{1}{Z(\theta^*)} \approx \frac{1}{M} \sum_{m=1}^{M} \frac{q_u(u^{(m)}|\theta^*, y)}{\gamma(u^{(m)}|\theta^*)}.$$

Noisy methods

- The use of “inexact approximate” or “noisy” methods in which an exact method is approximated **without** resulting in the correct target distribution.
- Focus on doubly intractable problems
  - strong link to work on other types of intractable likelihood.
- In particular, that an exact sampler does not exist for $u^* \sim f(.|\theta^*)$.
- Alternatives:
  - Russian roulette (Lyne et al., 2015);
  - use a long run of an MCMC in place of an exact sampler (Caimo and Friel, 2011; Everitt, 2012).

Error of estimates: noisy IS

- Noisy importance sampling and sequential Monte Carlo: Everitt et al (2016).
- Under some simplifying assumptions, noisy importance sampling is more efficient (in terms of mean squared error) compared to an exact-approximate algorithm if

$$\frac{1}{P}\left(\mathrm{Var}_q\left[w(\theta)+b(\theta)\right]+\mathbb{E}_q[\grave{\sigma}_\theta^2]\right)+\mathbb{E}_q[b(\theta)]^2 < \frac{1}{P}\left(\mathrm{Var}_q\left[w(\theta)\right]+\mathbb{E}_q[\acute{\sigma}_\theta^2]\right),$$

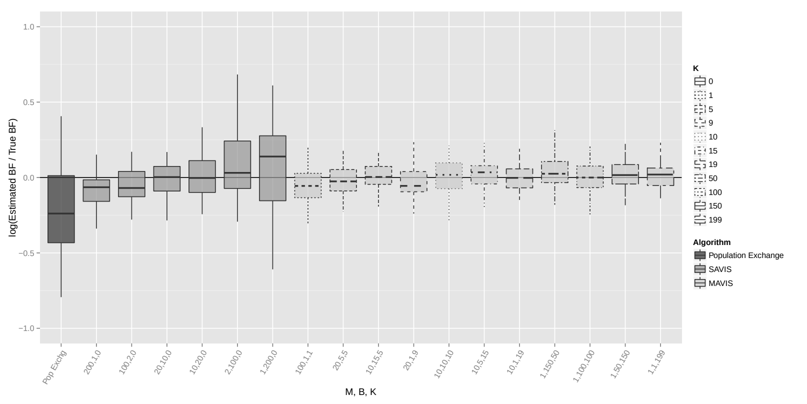
where $b(\theta)>0$ is the bias of the noisy weights, $\grave{\sigma}_\theta^2$ is the variance of the noisy weights, $\acute{\sigma}_\theta^2$ is the variance of the exact-approximate weights and

$$w(\theta) := \frac{p(\theta)\gamma(y|\theta)}{Z(\theta)q(\theta)}.$$

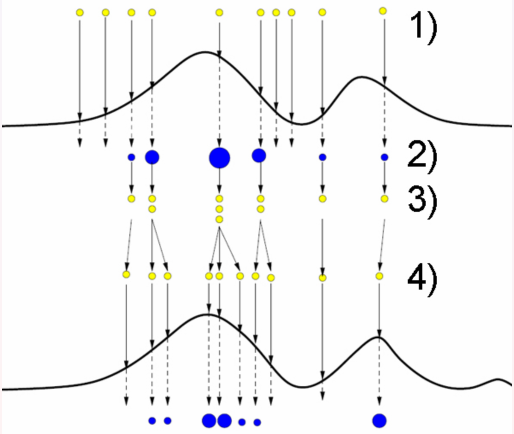
Application to Ising models

- An Ising model is a pairwise Markov random field with binary variables.
- Reanalyse the data from Friel (2013), which consists of 20 realisations from a first-order 10×10 Ising model and 20 realisations from a second-order 10×10 Ising model.
- Compare
  - population exchange;
  - SAVIS / MAVIS

Ising models: results



Sequential Monte Carlo



SMC samplers

An iteration of an SMC algorithm at target $t+1$.

- For $p = 1 : P$
  - Update $\theta_t^{(p)}$ to $\theta_{t+1}^{(p)}$ using some kernel K.
- For $p = 1 : P$
  - Reweight: find $\widetilde{w}_{t+1}^{(p)}$, so that the $\left(\theta_{t+1}^{(p)}, \widetilde{w}_{t+1}^{(p)}\right)$ are (unnormalised) weighted points from $p_{t+1}(.|y)$.
- Normalise $\left\{\widetilde{w}_{t+1}^{(p)}\right\}_{p=1}^{P}$ to give $\left\{w_{t+1}^{(p)}\right\}_{p=1}^{P}$.
- Resample the weighted points if some threshold is met.

- An estimate of the marginal likelihood is given by $\prod_{t=1}^{T} \sum_{p=1}^{P} \widetilde{w}_t^{(p)}$.

SMC for doubly intractable models

- Suppose that our sequence of distributions is

$$\pi_t(\theta|y) = p(\theta)f_t(y|\theta) = p(\theta)\frac{\gamma_t(y|\theta)}{Z_t(\theta)}.$$

- Using an MCMC kernel, we obtain an weight of

$$\widetilde{w}_t^{(p)} = \frac{\gamma_t(y|\theta_{t-1}^{(p)})}{\gamma_{t-1}(y|\theta_{t-1}^{(p)})}\frac{Z_{t-1}(\theta_{t-1}^{(p)})}{Z_t(\theta_{t-1}^{(p)})}w_{t-1}^{(p)}. \quad (3)$$

SMC for doubly intractable models

- Use an unbiased IS estimator of the ratio of Zs

$$\widehat{\frac{Z_{t-1}(\theta_{t-1}^{(p)})}{Z_t(\theta_{t-1}^{(p)})}} = \frac{1}{M} \sum_{m=1}^{M} \frac{\gamma_{t-1}(u_t^{(m,p)} | \theta_{t-1}^{(p)})}{\gamma_t(u_t^{(m,p)} | \theta_{t-1}^{(p)})}, \tag{4}$$

  where $u_t^{(p,m)} \sim f_t(.|\theta_{t-1}^{(p)})$.

- Viewed on an extended space, this is not *quite* the SMC construction of Del Moral et al. (2006), but is still exact.

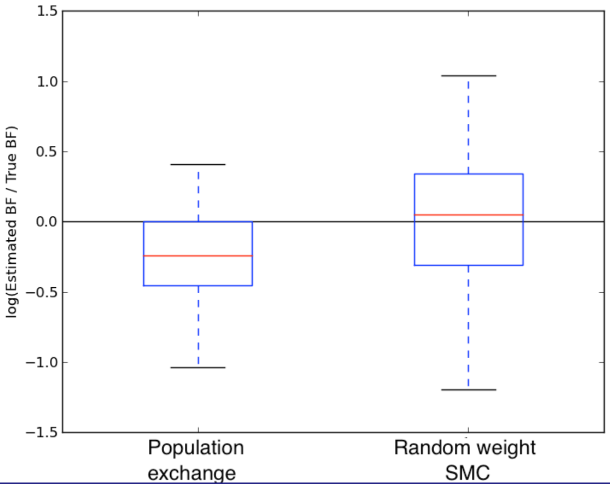
Sequence of distributions

- Suppose there are T data points.
- Use $\pi_t(\theta|y) = p(\theta)f_t(y|\theta)$ with

$$f_t(y|\theta) = f\left(y_{1:t/T}|\theta\right) = \gamma(y_{1:t/T}|\theta)/Z_t(\theta), \tag{5}$$

  i.e. essentially we add in one data point for each increment of t.
- As in Chopin 2002, or Chopin et al. 2013.
- Then it is simple to use IS to estimate $1/Z_t(\theta)$ (and ratios of Zs).

Sequential Monte Carlo results



An alternative choice

- Why not use $\pi_t(\theta|y) = p(\theta) f_t(y|\theta)$ with
$$f_t(y|\theta) = f^{t/T}(y|\theta)?$$
- Suppose unbiased estimates $\hat{f}$ of f are available
  - includes doubly intractable situation, but more general than this.
- Can we use
$$f_t(y|\theta) = \hat{f}^{t/T}(y|\theta)?$$
- Results in biased estimates of the weights
  - noisy SMC.

An alternative choice

- Why not use $\pi_t(\theta|y) = p(\theta) f_t(y|\theta)$ with

$$f_t(y|\theta) = f^{t/T}(y|\theta)?$$

- Suppose unbiased estimates $\hat{f}$ of f are available
  - includes doubly intractable situation, but more general than this.
- Can we use

$$f_t(y|\theta) = \hat{f}^{t/T}(y|\theta)?$$

- Results in biased estimates of the weights
  - noisy SMC.

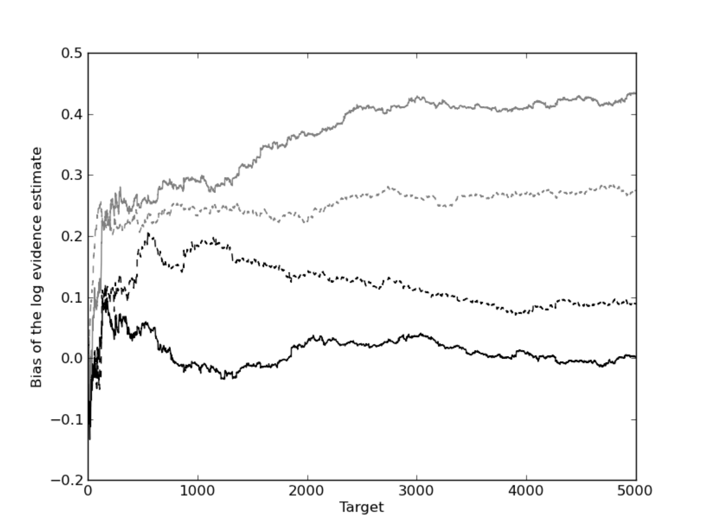
An alternative choice

- Why not use $\pi_t(\theta|y) = p(\theta)f_t(y|\theta)$ with
$$f_t(y|\theta) = f^{t/T}(y|\theta)?$$
- Suppose unbiased estimates $\hat{f}$ of f are available
  - includes doubly intractable situation, but more general than this.
- Can we use
$$f_t(y|\theta) = \hat{f}^{t/T}(y|\theta)?$$
- Results in biased estimates of the weights
  - noisy SMC.

Noisy SMC: strong mixing assumptions

- In Everitt et al (2016), we
  - use biased weights at every step of the SMC;
  - are interested in how the error accumulates as the SMC algorithm iterates.
- Under
  - strong mixing assumptions (stronger than a global Doeblin condition)
  - a small difference between exact and noisy weight functions
- Obtain a uniform bound on total-variation discrepancy between the iterated target distributions of the exact and noisy methods
  - strong mixing can prevent the accumulation of error even in systems with biased weights.

Noisy SMC: empirical results



Marginal SMC

- Marginal SMC (very similar to PMC) offers a solution
  - integrates over the previous target, rather than sampling from the path space of targets
  - thus bias does not accumulate
  - has the correct target as long as $\hat{f}$ is unbiased.
- Weight update is

$$\tilde{w}_t^{(p)} = \frac{p\left(\theta_t^{(p)}\right)\hat{f}^{t/T}\left(y|\theta_t^{(p)}\right)}{\sum_{r=1}^{P} w_{t-1}^{(r)} K_t\left(\theta_t^{(p)} \mid \theta_{t-1}^{(r)}\right)}.$$

- Can be used very generally with estimated likelihoods.

Marginal SMC

- Adaptation is natural.
- $\hat{f}$ computed at early stages of the SMC can be used in the later stages.
- Population of points moves from a disperse distribution to a concentrated one
  - when using pre-computation, helps avoid the problem of having poor estimates in regions that have not been visited (e.g. the tails).
- Avoids stickiness of MCMC chain caused to high variance estimates.

SAV revisited

- Suppose we alter the (unnormalised weight) to be

$$w^{(p)} = \frac{p(\theta^{(p)})\gamma(y|\theta^{(p)})}{q(\theta^{(p)})} \frac{Z(\tilde{\theta})}{Z(\theta^{(p)})},$$

for some $\tilde{\theta}$.

- We now require an estimate of

$$\frac{Z(\tilde{\theta})}{Z(\theta^{(p)})}.$$

- Now

$$\widehat{\frac{Z(\tilde{\theta})}{Z(\theta^{(p)})}} = \frac{\gamma\left(u|\tilde{\theta}\right)}{\gamma\left(u|\theta^{(p)}\right)}$$

with $u \sim f\left(\cdot|\theta^{(p)}\right)$. Use $\hat{f}^{t/T}$ within marginal SMC.

Low variance estimates

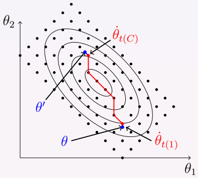
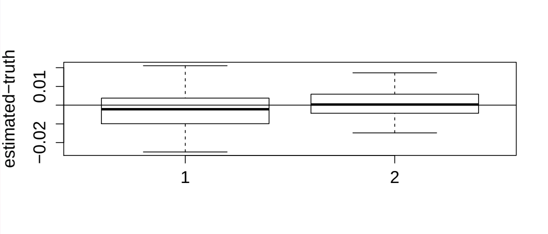


Image from Nial Friel.

$$\frac{\widehat{Z(\tilde{\theta})}}{Z(\theta^{(p)})} = \frac{\widehat{Z(\theta_1)}}{Z(\theta^{(p)})} \times \frac{\widehat{Z(\theta_2)}}{Z(\theta_1)} \times \dots \times \frac{\widehat{Z(\tilde{\theta})}}{Z(\theta_m)}$$

- Here
  - $\tilde{\theta} = \frac{1}{P} \sum_{p=1}^{P} \theta_{t-1}^{(p)}$;
  - $\theta_1, \dots, \theta_m$ are a path of previously visited values from previous steps of the SMC.

Application to precision estimation



- Estimating the posterior expectation of θ for a 10×10 Ising model.
- Marginal SMC with 50 particles and 20 targets (1: without path; 2: with path).
- Compare to a long run of the exchange algorithm.

Synthetic likelihood

- From Wood (2010), use the estimate

$$\widehat{f}_{\mathsf{SL}}\left(S(y)|\theta\right) = \mathcal{N}\left(S(y);\widehat{\mu}_{\theta},\widehat{\Sigma}_{\theta}\right),$$

  where

$$\widehat{\mu}_{\theta} = \frac{1}{M}\sum_{m=1}^{M} S\left(u^{(m)}\right),$$

$$\widehat{\Sigma}_{\theta} = \frac{ss^{T}}{M-1},$$

  for $\{u^{(m)}\}_{m=1}^{M} \sim f(\cdot|\theta^{*})$.
- A type of noisy MCMC.

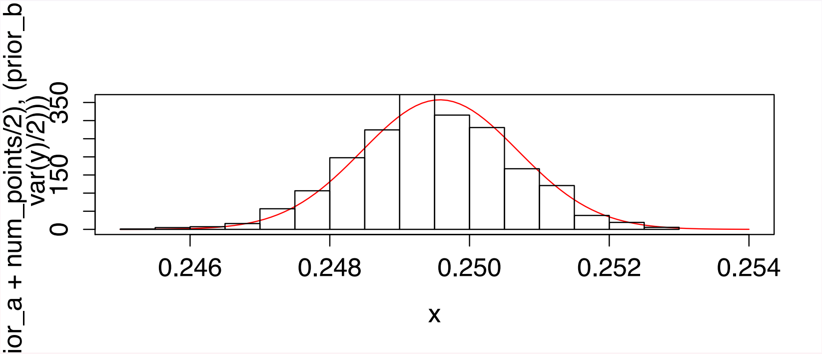
Regression idea

- If we are prepared to accept a little bias...
- ... wasteful to estimate $\widehat{f}(y|\theta)$ independently for each theta.
- We could try to exploit local smoothness of f in θ by estimating a regression of f on θ.
- Use the regression predictions as the likelihood
  - introduces a bias;
  - lower variance;
  - also explored in other papers...

Subsampling

- **Problem:** expensive if the dimension N of y is large.
- **Approach:** estimate regression of μ_θ and Σ_θ on θ via estimates based on subsamples of y (and using a small M) and use the regression predictions
  - reduces the variance of these estimates;
  - also see Moores et al. (2015) (regression without subsampling).
- Use within marginal SMC.
  - where $f_t(y|\theta) = \widehat{f}_{\text{SL}}^{t/T}(S(y)|\theta)$.

Application to precision estimation



True data size: $N = 100,000$.

Size of data simulated each time: 1,000.

Simulations per iteration: $M = 10$.

Conclusions

- Use exact methods where possible...
- ... however the bias from a noisy method may be small compared to errors resulting from commonly accepted approximate techniques such as ABC (and also the Monte Carlo variance).
- What is the best we can do fo some finite computational budget?
- Marginal SMC is useful when working with estimated likelihoods
  - many potential applications.

Conclusions

- Use exact methods where possible...
- ... however the bias from a noisy method may be small compared to errors resulting from commonly accepted approximate techniques such as ABC (and also the Monte Carlo variance).
- What is the best we can do fo some finite computational budget?
- Marginal SMC is useful when working with estimated likelihoods
  - many potential applications.

Conclusions

- Use exact methods where possible...
- ... however the bias from a noisy method may be small compared to errors resulting from commonly accepted approximate techniques such as ABC (and also the Monte Carlo variance).
- What is the best we can do fo some finite computational budget?
- Marginal SMC is useful when working with estimated likelihoods
  - many potential applications.