Importance sampling type estimators based on approximate marginal MCMC

Matti Vihola
Department of Mathematics and Statistics, University of Jyväskylä

Bayesian Computation for High-Dimensional Statistical Models, Singapore, 31 Aug 2018

Joint work with: Jouni Helske (Linköping), Jordan Franks (Jyväskylä),
Ajay Jasra (Singapore) and Kody Law (Manchester)

Overview

Bayesian latent variable models

Some popular inference algorithms

Importance sampling type estimator based on marginal MCMC

Examples

Discussion

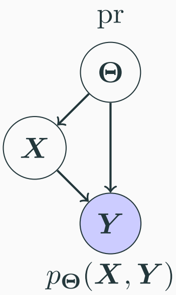
Bayesian latent variable models

Bayesian latent variable models

The following abstraction applies to many relevant statistical models:

- $\boldsymbol{\Theta}$ is a vector of (hyper)parameters
- $\boldsymbol{X}$ is a vector of latent variables
- $\boldsymbol{Y}$ is a vector of observations

Only $\boldsymbol{Y} = \boldsymbol{y}^*$ observed, both $\boldsymbol{\Theta}$ and $\boldsymbol{X}$ are unknown.



The model is defined in terms of the following conditonal laws:

- $\boldsymbol{\Theta} \sim \mathrm{pr}(\,\cdot\,)$.
- $(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{\Theta} \sim p_{\boldsymbol{\Theta}}(\,\cdot\,)$
  (Often, $p_{\boldsymbol{\Theta}}(\boldsymbol{X}, \boldsymbol{Y}) = f_{\boldsymbol{\Theta}}(\boldsymbol{X}) g_{\boldsymbol{\Theta}}(\boldsymbol{Y} \mid \boldsymbol{X})$, but this is not relevant here.)

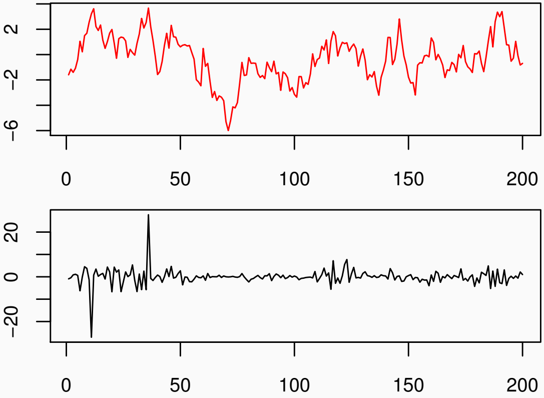
We are interested in the posterior of $(\boldsymbol{\Theta}, \boldsymbol{X})$ after observing $\boldsymbol{Y} = \boldsymbol{y}^*$:

$$\boxed{\pi(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}^*) \propto p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}^*) = \mathrm{pr}(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*).}$$

Example of $p_{\Theta}(\boldsymbol{X}, \boldsymbol{Y})$: Stochastic volatility model

- $\boldsymbol{\Theta} = (\phi, \sigma_x, \sigma_y)$
- $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(T)})$ stationary Gaussian AR(1) with parameters (ϕ, σ_x^2).
- The observations $\boldsymbol{Y} = (Y^{(1)}, \ldots, Y^{(T)})$ are zero-mean Gaussian with $\mathrm{sd}(Y^{(t)}) = \sigma_y \exp(X^{(t)})$,

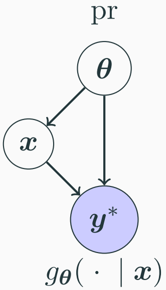
One realisation of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $\boldsymbol{\theta} = (0.9, 1, 2)$ and $T = 200$.



Challenges for inference

Typical scenario in a latent variable model:

- The hyperparameters $\boldsymbol{\Theta}$ are low-dimensional
    - $\dim(\boldsymbol{\Theta}) = 3$ in the SV-example.
- The latent variables $\boldsymbol{X}$ are high-dimensional
    - Often, $\dim(\boldsymbol{X}) \propto \dim(\boldsymbol{Y})$.
    - $\dim(\boldsymbol{X}) = 200$ in the SV-example.



Standard 'out-of-the-box' inference (e.g. using `BUGS`, `Stan`...):

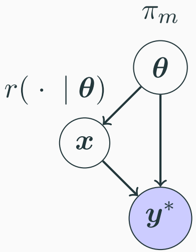
- Simulate MCMC chain $\boldsymbol{Z}_k = (\boldsymbol{\Theta}_k, \boldsymbol{X}_k)$, targetting π.
- High overall dimension & high correlations $\implies$ inefficient

Some popular inference algorithms

Structure of the model & notation for inference

Consider the following factorisation of the posterior:

$$\pi(\boldsymbol{\theta}, \boldsymbol{x}) = \pi_m(\boldsymbol{\theta}) r(\boldsymbol{x} \mid \boldsymbol{\theta}),$$



where the marginal posterior density and the corresponding conditional are given as follows:

$$\pi_m(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta}, \boldsymbol{x}) \mathrm{d}\boldsymbol{x} \propto \mathrm{pr}(\boldsymbol{\theta}) L(\boldsymbol{\theta})$$

$$r(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*)}{L(\boldsymbol{\theta})}$$

with the marginal likelihood L taking the form:

$$L(\boldsymbol{\theta}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*) \mathrm{d}\boldsymbol{x}.$$

Separate algorithms for parameters & latents?

- $\boldsymbol{\theta}$ low-dimensional, but π_m often non-standard
  $\implies$ Non-parametric approximation, such as MCMC
- Problems (unless $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*)$ is of specific form such as Gaussian):
  - $L(\boldsymbol{\theta})$ is intractable.
  - $r(\boldsymbol{x} \mid \boldsymbol{\theta})$ is intractable.
- Two (succesful branches of) solutions:
  - Approximate $L(\boldsymbol{\theta})$ and $r(\boldsymbol{x} \mid \boldsymbol{\theta})$ analytically.
  - Approximate $L(\boldsymbol{\theta})$ and $r(\boldsymbol{x} \mid \boldsymbol{\theta})$ using a specialised Monte Carlo algorithm.

Inference: Integrated nested Laplace approximation (INLA)

(Rue, Martino & Chopin, 2009, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*)

- Suppose $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) g_{\boldsymbol{\theta}}(\boldsymbol{y})$ where $f_{\boldsymbol{\theta}}(\,\cdot\,)$ is Gaussian (& $g_{\boldsymbol{\theta}}$ of certain form)
- For any given $\boldsymbol{\theta}$, use Gaussian (Laplace) approximation $\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*)$
  $\implies$ approximate likelihood $L_a(\boldsymbol{\theta}) = \int \hat{p}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*) \mathrm{d}\boldsymbol{x}$
- Take a finite number of points $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ and approximate the full posterior as (something like)

$$\hat{\pi}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\boldsymbol{x}) = \frac{\sum_{k=1}^n w_k \mathrm{pr}(\boldsymbol{\theta}_k) \hat{p}_{\boldsymbol{\theta}_k}(\boldsymbol{x}, \boldsymbol{y}^*) \delta_{\boldsymbol{\theta}_k}(\mathrm{d}\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}}{\sum_{j=1}^n w_j \mathrm{pr}(\boldsymbol{\theta}_j) L_a(\boldsymbol{\theta}_j)}.$$

  where the weight w_k depends on the strategy how $(\boldsymbol{\theta}_i)$ chosen...
  - (Further marginal corrections may be applied as well...)
- There is an approximation error (which does not vanish if $n \to \infty$)

(Andrieu, Doucet & Holenstein, 2010, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*)

- For any given $\boldsymbol{\theta}$, it is straightforward to generate random variables $(V^{(i)}, \boldsymbol{X}^{(i)})$, with $V^{(i)} \geq 0$, using particle filter (PF), which satisfy

$$\mathbb{E}\left[\sum_{i=1}^{m} V^{(i)}\right] = L(\boldsymbol{\theta}), \qquad \text{and} \qquad \mathbb{E}\left[\sum_{i=1}^{m} V^{(i)} f(\boldsymbol{X}^{(i)})\right] = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^*) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

- The algorithm:
  - Implement Metropolis-Hastings $(\boldsymbol{\Theta}_k)_{k \geq 1}$ targetting π_m, using $\sum_{i=1}^{m} V^{(i)}$ in place of $L(\boldsymbol{\theta})$.
  - Construct an approximation of full $\pi(\boldsymbol{\theta}, \boldsymbol{x})$ using $(V^{(i)}, \boldsymbol{X}^{(i)})$ above.

Particle marginal Metropolis-Hastings (PMMH) algorithm:

- Draw a new proposal $\tilde{\mathbf{\Theta}}_k \sim q(\mathbf{\Theta}_{k-1}, \,\cdot\,)$
- $\boxed{\text{Run PF with } \boldsymbol{\theta} = \tilde{\mathbf{\Theta}}_k} \longrightarrow (\tilde{V}_k^{(i)}, \tilde{\boldsymbol{X}}_k^{(i)})$.
- Accept and set $(\mathbf{\Theta}_k, V_k^{(i)}, \boldsymbol{X}_k^{(i)}) \leftarrow (\tilde{\mathbf{\Theta}}_k, \tilde{V}_k^{(i)}, \tilde{\boldsymbol{X}}_k^{(i)})$ With probability

$$\min\left\{1, \frac{\mathrm{pr}(\tilde{\mathbf{\Theta}}_k)\big(\sum_{i=1}^m \tilde{V}_k^{(i)}\big) q(\tilde{\mathbf{\Theta}}_k, \mathbf{\Theta}_{k-1})}{\mathrm{pr}(\mathbf{\Theta}_{k-1})\big(\sum_{i=1}^m V_{k-1}^{(i)}\big) q(\mathbf{\Theta}_{k-1}, \tilde{\mathbf{\Theta}}_k)}\right\};$$

otherwise reject and set $(\mathbf{\Theta}_k, V_k^{(i)}, \boldsymbol{X}_k^{(i)}) \leftarrow (\mathbf{\Theta}_{k-1}, V_{k-1}^{(i)}, \boldsymbol{X}_{k-1}^{(i)})$

What is nice about this is that:

- This is valid MCMC, in the sense that

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\sum_{i=1}^{m} V_k^{(i)} f(\boldsymbol{\Theta}_k, \boldsymbol{X}_k)}{\sum_{i=1}^{m} V_k^{(i)}} \xrightarrow{n\to\infty} \int f(\boldsymbol{\theta}, \boldsymbol{x})\pi(\boldsymbol{\theta}, \boldsymbol{x})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{x}, \qquad \text{(a.s.)}$$

  (under a minimal Harris recurrence assumption)
- $\implies$ PMCMC provides (asymptotically) exact inference (as $n \to \infty$).

(NB: There is no asymptotic in 'number of particles' m...
...but m must be 'large enough' to make the MCMC mix sufficiently well...)

Should I use INLA or PMCMC?

- (When applicable) INLA[1] is fast, and often very accurate
  - How accurate? How can you tell?
- PMCMC computationally demanding, but exact (asymptotically)
  - Does not require 'nearly Gaussian' structure of $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$.
  - Might need large m to work well $\implies$ slow
  - Might still be 'sticky' (slower than geometric if $\sum_{i=1}^{m} V^{(i)}$ are unbounded. . .)
- How about combining ideas both from INLA and PMCMC:

→ Monte Carlo correction/diagnosis for INLA output, or

→ Laplace approximations to speed up PMCMC. . .

[1] The same arguments hold with any other approximate scheme in place of INLA'!

Approximations for speeding up PMMH — delayed acceptance

(Christen & Fox, 2005, *J. Comput. Graph. Statist.*)

Trick to make MCMC faster by using an approximation-based 'screening'.

- Draw a new proposal $\tilde{\mathbf{\Theta}}_k \sim q(\mathbf{\Theta}_{k-1}, \,\cdot\,)$
- With probability

$$\min\left\{1, \frac{\mathrm{pr}(\tilde{\mathbf{\Theta}}_k) L_a(\tilde{\mathbf{\Theta}}_k) q(\tilde{\mathbf{\Theta}}_k, \mathbf{\Theta}_{k-1})}{\mathrm{pr}(\mathbf{\Theta}_{k-1}) L_a(\mathbf{\Theta}_{k-1}) q(\mathbf{\Theta}_{k-1}, \tilde{\mathbf{\Theta}}_k)}\right\}$$

  continue to the next step, otherwise reject.
- $\boxed{\text{Run PF with } \boldsymbol{\theta} = \tilde{\mathbf{\Theta}}_k} \longrightarrow (\tilde{V}_k^{(i)}, \tilde{\boldsymbol{X}}_k^{(i)})$
- With probability

$$\min\left\{1, \frac{\left(\sum_{i=1}^m \tilde{V}_k^{(i)}\right) / L_a(\tilde{\mathbf{\Theta}}_k)}{\left(\sum_{i=1}^m V_{k-1}^{(i)}\right) / L_a(\mathbf{\Theta}_{k-1})}\right\}$$

  accept, otherwise reject.

Importance sampling type estimator based on marginal MCMC

Approximations for speeding up PMMH — importance sampling type

(Review, consistency and CLTs: V, Helske, Franks, arXiv:1609.02541)

Phase 1: MCMC which targets the approximate marginal $\pi_a(\boldsymbol{\theta}) \propto \mathrm{pr}(\boldsymbol{\theta} L_a(\boldsymbol{\theta})$

- Draw a new proposal $\tilde{\mathbf{\Theta}}_k \sim q(\mathbf{\Theta}_{k-1}, \,\cdot\,)$
- With probability

$$\min\left\{1, \frac{\mathrm{pr}(\tilde{\mathbf{\Theta}}_k)L_a(\tilde{\mathbf{\Theta}}_k)q(\tilde{\mathbf{\Theta}}_k, \mathbf{\Theta}_{k-1})}{\mathrm{pr}(\mathbf{\Theta}_{k-1})L_a(\mathbf{\Theta}_{k-1})q(\mathbf{\Theta}_{k-1}, \tilde{\mathbf{\Theta}}_k)}\right\}$$

accept $\mathbf{\Theta}_k = \tilde{\mathbf{\Theta}}_k$; otherwise reject $\mathbf{\Theta}_k = \mathbf{\Theta}_{k-1}$.

Phase 2: For $k = 1, \ldots, n$, $\boxed{\text{run PF with } \boldsymbol{\theta} = \mathbf{\Theta}_k} \longrightarrow (V_k^{(i)}, X_k^{(i)})$ and calculate

$$E_n = \frac{\sum_{k=1}^n \sum_{i=1}^m W_k^{(i)} f(\mathbf{\Theta}_k, \boldsymbol{X}_k^{(i)})}{\sum_{j=1}^n \sum_{\ell=1}^m W_j^{(\ell)}} \qquad \text{where} \qquad W_k^{(i)} = \frac{V_k^{(i)}}{L_a(\mathbf{\Theta}_k)}.$$

Why IS might be better than DA?

- Phase 2 corrections entirely independent ('post-processing') $\Longrightarrow$ parallelisable $\Longrightarrow$ scalable.
- Allows for calculating the correction only for accepted states ('jump chain') $\Longrightarrow$ less expensive than DA
- Allows for (further) *thinning* before (expensive) correction $\Longrightarrow$ further savings
- The approximate marginal MCMC $(\boldsymbol{\Theta}_k)$ need not rely on estimators $\Longrightarrow$ safer & easier to implement efficiently (e.g. adaptive MCMC...)
- The MCMC $(\boldsymbol{\Theta}_k)$ need not be reversible $\Longrightarrow$ new exciting non-reversible samplers readily applicable!
- Non-negativity of the estimator W_k not required $\Longrightarrow$ allows for direct 'debiasing' tricks (or 'randomised multi-level Monte Carlo') (cf. Rhee & Glynn, *Oper. Res.* 2015; V, *Oper. Res.*, 2018)

General setup & assumptions

General setup for IS *type* estimators based on approximate marginal MCMC:

- $\pi(\boldsymbol{\theta}, \boldsymbol{x}) = \pi_m(\boldsymbol{\theta}) r(\boldsymbol{x} \mid \boldsymbol{\theta})$.
- $\pi_m \ll \pi_a$
- $(\boldsymbol{\Theta}_k)_{k \geq 1}$ MCMC Harris ergodic wrt π_a
- $\big(\xi_k\big)_{k \geq 1}$ conditionally independent finite random signed measures given $(\boldsymbol{\Theta}_k)_{k \geq 1}$, which form "proper weighting":

General	The LVM example
$\pi_m(\boldsymbol{\theta})$	$\propto \mathrm{pr}(\boldsymbol{\theta}) L(\boldsymbol{\theta})$
$\pi_a(\boldsymbol{\theta})$	$\propto \mathrm{pr}(\boldsymbol{\theta}) L_a(\boldsymbol{\theta})$
$\xi_k(f)$	$\sum_{i=1}^m W_k^{(i)} f(\boldsymbol{\Theta}_k, \boldsymbol{X}_k^{(i)})$

$$\mathbb{E}[\xi_k(\mathbf{1}) \mid \boldsymbol{\Theta}_k = \boldsymbol{\theta}] = w_u(\boldsymbol{\theta}), \qquad \text{where} \qquad w_u(\boldsymbol{\theta}) = c_w \frac{\pi_m(\boldsymbol{\theta})}{\pi_a(\boldsymbol{\theta})}, \quad c_w > 0$$

$$\mathbb{E}[\xi_k(f) \mid \boldsymbol{\Theta}_k = \boldsymbol{\theta}] = w_u(\boldsymbol{\theta}) \int r(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}, \boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

Consistency & CLT

- If $\pi_a(m^{(1)}) < \infty$ where $m^{(1)}(\boldsymbol{\theta}) = \mathbb{E}\big[|\xi_k(1)| + |\xi_k(f)| \;\big|\; \boldsymbol{\Theta}_k = \boldsymbol{\theta}\big]$, then

$$E_n = \frac{\sum_{k=1}^n \xi_k(f)}{\sum_{j=1}^n \xi_j(\mathbf{1})} \xrightarrow[\text{a.s.}]{n\to\infty} \pi(f) = \int f(\boldsymbol{\theta}, \boldsymbol{x})\pi(\boldsymbol{\theta}, \boldsymbol{x})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{x}.$$

- Suppose further that (for instance):
  - $\pi_a(m^{(2)}) < \infty$ with $m^{(2)} = \mathbb{E}\big[\xi_k(\bar{f})^2 \;\big|\; \boldsymbol{\Theta}_k = \boldsymbol{\theta}\big]$ where $\bar{f}(\boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{\theta}, \boldsymbol{x}) - \pi(f)$,
  - $(\boldsymbol{\Theta}_k)_{k\geq 1}$ follows P which is aperiodic and reversible, with asymptotic variance $\mathrm{Var}(w_u\bar{f}^*, P) < \infty$, where $\bar{f}^*(\boldsymbol{\theta}, \boldsymbol{x}) = \int \bar{f}(\boldsymbol{\theta}, \boldsymbol{x}')r(\boldsymbol{x}' \mid \boldsymbol{\theta})\mathrm{d}\boldsymbol{x}'$,

Then,

$$\sqrt{n}\big[E_n - \pi(f)\big] \xrightarrow[\text{d}]{n\to\infty} N\Bigg(0, \overbrace{\frac{\mathrm{Var}(w_u\bar{f}^*, P)}{c_w^2}}^{\text{MCMC}} + \overbrace{\frac{\pi_a(v)}{c_w^2}}^{\text{IS corr}}\Bigg),$$

where $v(\boldsymbol{\theta}) = \mathrm{Var}\big(\xi_k(\bar{f}) \;\big|\; \boldsymbol{\Theta}_k = \boldsymbol{\theta}\big)$.

Pseudo-marginal approximate chain

- The IS-type correction may be applied also when $(\mathbf{\Theta}_k, U_k)_{k\geq 1}$ is a pseudo-marginal chain arising from estimators $\tilde{U}_{\boldsymbol{\theta}}$ satisfying $E[\tilde{U}_{\boldsymbol{\theta}}] = L_a(\boldsymbol{\theta})$.

In the pseudo-marginal case, consistency is more delicate:

- If $\tilde{U}_{\boldsymbol{\theta}} > 0$ a.s., then then we may always use $W_k^{(i)} = V_k^{(i)}/U_k$.
- When $\mathbb{P}(\tilde{U}_{\boldsymbol{\theta}} = 0)$ depends on $\boldsymbol{\theta}$, this must be accounted for.
- For instance, $V_k^{(i)}$ is constructed independent of U_k, then we must compensate for an extra factor $p(\boldsymbol{\theta}) = \mathbb{P}(U_{\boldsymbol{\theta}} > 0)\ldots \rightsquigarrow$ lazy ABC

Examples

State space model with linear-Gaussian state dynamics

- State dynamics linear-Gaussian.
- Family of non-linear/non-Gaussian observation models.
- Approximate inference based on Laplace approximation (Durbin & Koopman, *Biometrika*, 1997).
- Proper weighting based on:
  (i) Bootstrap particle filter (BSF).
  (ii) Simple importance sampling & antithetic variables (SPDK, Shephard & Pitt, *Biometrika*, 1997; Durbin & Koopman, *Biometrika*, 1997)
  (iii) ψ-auxiliary particle filter (ψ-APF: bootstrap PF for 'Laplace twisted model;' see Guarniero, Johansen & Lee, *JASA*, 2017)
- Simple IS-corrected estimator (IS1) or estimator based on jump chain (IS2).
- Compare against direct pseudo-marginal (PM) and delayed acceptance (DA).

State space model with linear-Gaussian state dynamics: empirical results

(Stochastic volatility model with $T = 5473$ observations, S&P index data. The numbers are 'inverse relative efficiencies' (avg. time (h) $\times$ MSE) — lower is better.)

	BSF		SPDK				ψ-APF					
	AI	AI[G]	IS2	IS2[8]	PM	DA	IS1	IS2	PM	DA	IS1	IS2
Time	1.3	0.2	25.2	4.6	4.4	1.9	2.8	1.5	2.4	1.4	1.5	1.3
ϕ	0.083	0.062	0.304	0.050	1.015	0.696	0.684	0.483	0.021	0.024	0.009	0.017
σ_η	0.726	0.298	0.483	0.096	3.090	3.307	0.603	0.710	0.044	0.055	0.016	0.028
ν	0.008	0.747	0.287	0.042	1.208	2.544	0.228	0.404	0.026	0.027	0.010	0.020
X_1	0.133	0.035	0.321	0.071	3.054	1.883	0.346	0.373	0.029	0.026	0.007	0.018
X_{5473}	1.887	0.417	0.540	0.112	6.574	1.871	0.444	0.810	0.057	0.064	0.012	0.039

Discretely observed (time-discretised) diffusion

- ‘Ideal’ state dynamics follows a stochastic differential equation (SDE).
    - Cannot simulate exactly from the ideal transition.
    - Easy to simulate from time-discretised model (Euler, Milstein, . . .).
    - The denser discretisation, the more simulation costs.
- Conditionally independent observations at discrete times.
- Approximate inference: particle marginal Metropolis-Hastings (PMMH) with ‘coarse’ (and cheap) time-discretisation.
- Correction with particle filter using ‘fine’ time-discretisation.

Discretely observed (time-discretised) diffusion: empirical results

(Geometric Brownian motion observed at integer times, linear-Gaussian observations of log-state, Milstein discretisation. Parallel implementation with 48 cores, time mins.)

	Mean						IRE				
Init.		Prior mean			Prior sample		Prior mean			Prior sample	
	GT	DA	IS2	IS2$^{\mathrm{t}}$	DA	IS2	DA	IS2	IS2$^{\mathrm{t}}$	DA	IS2
Time	—	12.3	3.4	1.9	14.0	3.3	12.3	3.4	1.9	14.0	3.3
ν	0.053	0.061	0.053	0.053	0.064	0.053	0.069	0.004	0.002	0.135	0.004
σ_x	0.253	0.278	0.253	0.253	0.251	0.252	0.576	0.029	0.019	0.336	0.022
σ_y	1.058	1.054	1.058	1.058	1.083	1.058	0.088	0.020	0.014	1.010	0.022
X_1	1.254	1.273	1.254	1.246	1.243	1.252	0.670	0.109	0.119	0.805	0.103
X_{50}	2.960	2.953	2.966	2.935	20.773	2.971	12.605	1.880	2.074	4×10^6	2.308

Discretely observed (time-discretised) diffusion: empirical results

(Geometric Brownian motion observed at integer times, linear-Gaussian observations of log-state, Milstein discretisation. Parallel implementation with 48 cores, time mins.)

	Mean						IRE				
Init.		Prior mean			Prior sample		Prior mean			Prior sample	
	GT	DA	IS2	IS2^{t}	DA	IS2	DA	IS2	IS2^{t}	DA	IS2
Time	—	12.3	3.4	1.9	14.0	3.3	12.3	3.4	1.9	14.0	3.3
ν	0.053	0.061	0.053	0.053	0.064	0.053	0.069	0.004	0.002	0.135	0.004
σ_x	0.253	0.278	0.253	0.253	0.251	0.252	0.576	0.029	0.019	0.336	0.022
σ_y	1.058	1.054	1.058	1.058	1.083	1.058	0.088	0.020	0.014	1.010	0.022
X_1	1.254	1.273	1.254	1.246	1.243	1.252	0.670	0.109	0.119	0.805	0.103
X_{50}	2.960	2.953	2.966	2.935	20.773	2.971	12.605	1.880	2.074	4×10^6	2.308

Discretely observed diffusion: Ideal diffusion inference with randomised MLMC

- Instead of correcting with 'fine' dynamics, it is possible to do IS correction for ideal SDE dynamics (Franks, Jasra, Law & V, arXiv:1807.10259, 2018).
- The correction is based on
  - Debiasing trick/randomised MLMC (Rhee & Glynn, *Oper. Res.*, 2015) with
  - 'Δ-PF' (Jasra, Kamatani, Law & Zhou. *SIAM J. Sci. Comp.*, 2018).
- Detailed presentation in the closing workshop!

Discussion

Is IS correction always better than DA (with same n)?

Is IS correction always better than DA (with same n)?

No.

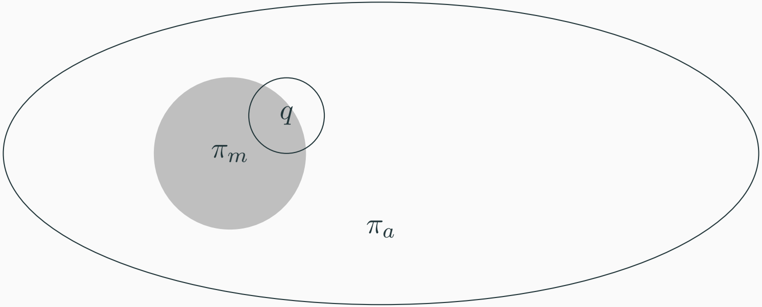
Is IS correction always better than DA (with same n)?

No.

More details Wed 12 Sep at 4pm...

DA can be much better than IS

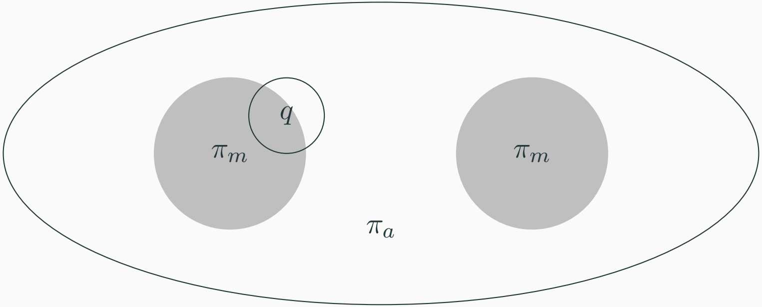
(Franks, V: arXiv:1706.09873)



DA better than IS: π_m and π_a are uniform, q uniform random walk. Approximate chain spends a lot of time outside the support of π_m.

But DA can also be much worse than IS...

(Franks, V: arXiv:1706.09873)



IS better than DA chain, which is reducible (cannot switch mode of π_m).

Can we say something about IS vs DA?

- In practice, we have $\pi_a \approx \pi_m$, which is clearly not the case in the examples above.
- Empirical results suggest that IS often improves on DA slightly

(Franks, V: arXiv:1706.09873):

- If $c_w^{-1} W_k \leq C$ a.s., then

$$\mathrm{Var}(\mathsf{IS}) \leq C\mathrm{Var}(\mathsf{DA}) + \bar{\pi}(\xi^2[C - c_w^{-1}W])$$

  where $\bar{\pi}$ corresponds to the stationary distribution of the DA chain.
  $\rightsquigarrow$ With parallelisation, IS might be a better choice...

- NB: In the LVM setting we may modify the likelihood approximation:
  - $L_a(\boldsymbol{\theta}) \to L_a(\boldsymbol{\theta}) + \epsilon$

  This leads to bounded weights if the likelihood estimators are bounded.

Concluding remarks

- If there is an approximation available, use it!
- IS type correction is a natural way to use the approximation
  - May be a useful alternative to DA pseudo-marginal algorithm (because of the several possible advantages)...
  - ... but not guaranteed to be uniformly better
- Our contributions:
  - arXiv:1609.02541: Review, consistency/CLT results; application in the state-space context, using Laplace approximation and coarse discretisation of diffusion model
  - arXiv:1706.09873: Theoretical bounds relating the efficiencies of IS/DA
  - arXiv:1807.10259: Full inference of SDE driven HMM based on randomised MLMC.
- Ongoing work:
  - Application beyond the state-space context.
  - Some insights for ABC-MCMC...

References

- **C. Andrieu, A. Doucet and R. Holenstein.**
  Particle Markov chain Monte Carlo methods.
  J. R. Stat. Soc. Ser. B. Stat. Methodol. 72(3), 269–342, 2010.
- **J. A. Christen and C. Fox.**
  Markov chain Monte Carlo using an approximation.
  J. Comput. Graph. Statist. 14(4), 795-810, 2005.
- **J. Franks, A. Jasra, K. Law and M. Vihola.**
  Unbiased inference for discretely observed hidden Markov model diffusions
  arXiv:1807.10259, 2018.
- **J. Franks and M. Vihola.**
  Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance
  arXiv:1706.09873, 2017.
- **M. Vihola, J. Helske and J. Franks.**
  Importance sampling type estimators based on approximate marginal MCMC.
  arXiv:1609.02541, 2016.

Thank you!