Sequential Monte Carlo with Derivative Information

Leah F South, A N Pettitt, N Friel and C C Drovandi

Sequential Monte Carlo

Sequential Monte Carlo (SMC) methods [1] can be powerful alternatives to Markov chain Monte Carlo (MCMC) methods [7] for performing inference on static Bayesian models. *SMC methods are adaptive, parallelisable and are more capable of dealing with multimodal or complex targets.*

Assume there is data $\mathbf{y}$ and interest is in a statistical model parameterised by $\boldsymbol{\theta}$. Likelihood annealing SMC traverses a population of N particles through a sequence of distributions defined by the power posteriors

$$\pi_t(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})^{\gamma_t}\pi(\boldsymbol{\theta}),$$

where $0 = \gamma_0 < \gamma_t < \gamma_T = 1$ and $0 < t < T$. A weighted particle set targeting π_t is represented by $\{W_t^i, \boldsymbol{\theta}_t^i\}_{i=1}^N$ where W_t^i is a normalised weight.

The effect of the likelihood is introduced smoothly through the following steps:

- Reweighting the particle set to target π_{t+1}. The new unnormalised weights are

$$w_{t+1}^i = W_t^i f(\mathbf{y}|\boldsymbol{\theta}_t^i)^{\gamma_{t+1}-\gamma_t}, \text{ for } i = 1, \ldots, N.$$

- Resampling.
- Diversifying the particles, often via several iterations of an MCMC kernel with a multivariate normal random walk (RW) proposal.

Evidence Estimation in SMC

The log evidence can be estimated using the stepping stone (SS) identity

$$\widehat{\log Z} = \sum_{t=1}^{T} \log \mathbb{E}_{\pi_{t-1}(\boldsymbol{\theta}|\mathbf{y})}[f(\mathbf{y}|\boldsymbol{\theta})^{\gamma_t-\gamma_{t-1}}], \quad (1)$$

or with the thermodynamic identity (TI)

$$\log Z = \int_0^1 \mathbb{E}_{\pi_t}[\log f(\mathbf{y}|\boldsymbol{\theta})]d\gamma, \quad (2)$$

which gives the log evidence as an integral with respect to the temperature γ [8]. We use a 2nd order quadrature approximation [2] for the integral in (2).

Using Derivative Information

The motivation for using the derivatives $\nabla_\theta \log \pi_t(\boldsymbol{\theta}|\boldsymbol{y})$ is that we would like to achieve the same level of precision with fewer likelihood evaluations.

Choice of Move Kernel

Using efficient move kernels leads to a higher acceptance rate and therefore fewer log likelihood evaluations.

Metropolis-adjusted Langevin algorithm (MALA, [3]) is an alternative to the popular RW proposal and it uses $\nabla_\theta \log \pi_t(\boldsymbol{\theta}|\boldsymbol{y})$:

$$q^{\phi_t}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_t^i) = \mathcal{N}(\boldsymbol{\theta}^*; \boldsymbol{\theta}_t^i + \frac{h_t^2}{2}\hat{G}_{\theta,t}^{-1}\nabla_\theta \log \pi_t(\boldsymbol{\theta}|\boldsymbol{y}), h_t^2\hat{G}_{\theta,t}^{-1}),$$

where the term $\hat{G}_{\theta,t}$ is a local measure of curvature for the log posterior and is referred to as the metric tensor. *We learn h from the population of particles.* The results in this poster are based on using the empirical covariance matrix for $\hat{\Sigma}_t$ for $\hat{G}_t^{-1}$ but if the second derivatives, $\nabla_\theta^2 \log \pi_t(\boldsymbol{\theta}|\boldsymbol{y}) \in \mathbb{R}^{d\times d}$, are available, these can be used to compute the observed or expected Fisher-Rao metric tensor $\hat{G}_{\theta,t}$ at $\boldsymbol{\theta}_t^i$ (MMALA,[3]).

Post-hoc Adjustment

Control variates can be used to get lower variance estimators of expectations $\mathbb{E}_{\pi_t(\theta)}[\varphi(\boldsymbol{\theta})]$. The general framework for control variates is to determine an auxiliary function $\tilde{\varphi}(\boldsymbol{\theta}) = \varphi(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ such that $\mathbb{E}_{\pi_t}[\tilde{\varphi}(\boldsymbol{\theta})] = \mathbb{E}_{\pi_t}[\varphi(\boldsymbol{\theta})]$ and $\mathbb{V}_{\pi_t}[\tilde{\varphi}(\boldsymbol{\theta})] < \mathbb{V}_{\pi_t}[\varphi(\boldsymbol{\theta})]$, where $\mathbb{V}_{p(\theta)}$ denotes the variance with respect to target $p(\boldsymbol{\theta})$. This can be achieved by choosing some random variable which is correlated with $\varphi(\boldsymbol{\theta})$ and has a known expectation. Here we use zero-variance control variates (ZV-CV, [6]), which require only the derivative of the log target or some unbiased estimator of this quantity.

We apply ZV-CV to all expectations in (1) *and* (2).

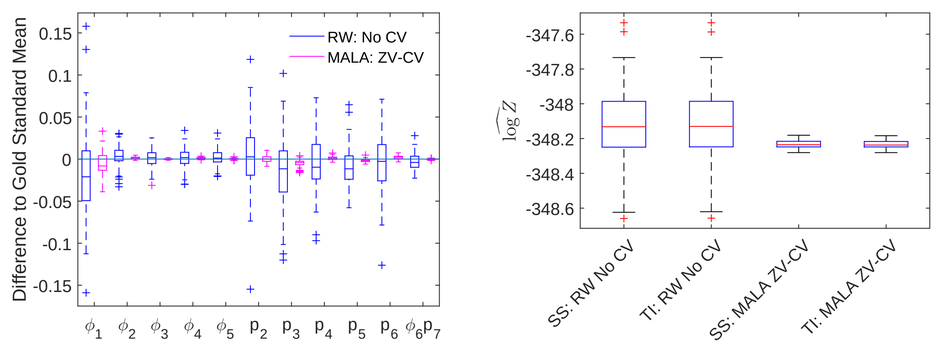
Recapture Example

Here we estimate the parameters of a Cormack-Jolly-Seber model based on the capture and recapture of a bird species [5]. The parameters are $\boldsymbol{\theta} = (\phi_1, \ldots, \phi_5, p_2, \ldots, p_6, \phi_6 p_7)$, where ϕ_i represents the probability of survival from year i to year $i+1$ and p_k represents the probability of being captured in year k. The likelihood for the model is

$$f(y|\theta) \propto \prod_{i=1}^{6}\left[1 - \sum_{k=i+1}^{7}\phi_i p_k \prod_{m=i+1}^{k-1}\phi_m(1-p_m)\right]^{D_i - \sum_{k=i+1}^{7} y_{ik}} \prod_{k=i+1}^{7}\left[\phi_i p_k \prod_{m=i+1}^{k-1}\phi_m(1-p_m)\right]^{y_{ik}},$$

where D_i is the number of birds released in year i and y_{ik} is the number of animals caught in year k out of the number released in year i. $\mathcal{U}(0,1)$ priors are used and all parameters are transformed to the real line for the move step.

Results: 100 SMC runs with $N = 1000$ particles are performed. *The figure below shows the improvement in posterior and evidence estimation that can be achieved with derivative information. RW uses between 1.5 and 25 times the number of log likelihood calculations that MALA uses.*



(a) Difference to the gold standard mean, on the transformed scale

(b) Log evidence estimates

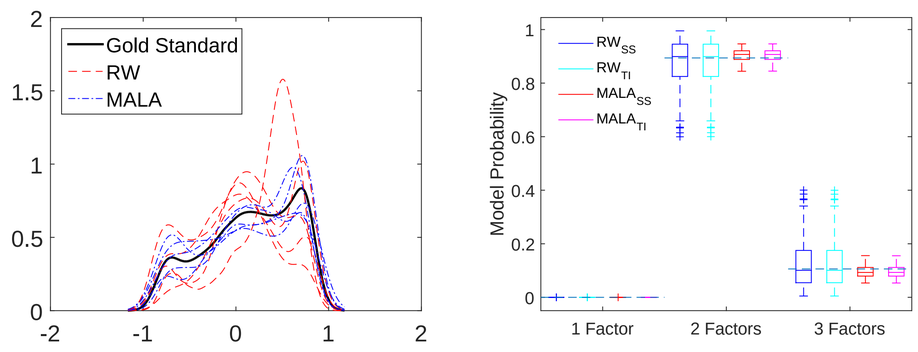
Figure: Performance with and without derivatives for posterior and log evidence estimation.

Factor Analysis Example

[4] use factor analysis models for data $\boldsymbol{Y}$ related to the exchange rate of 6 currencies relative to the British pound. The factor analysis models assume that $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$. To reduce model complexity, $\boldsymbol{\Omega}$ is parameterised by

$$\boldsymbol{\Omega} = \boldsymbol{\beta}\boldsymbol{\beta}^T + \boldsymbol{\Sigma},$$

where $\boldsymbol{\beta}$ is a $6 \times k$ lower triangular matrix with positive diagonal elements and $\boldsymbol{\Sigma}$ is a 6×6 diagonal matrix with positive elements. Here k is the number of factors. The prior and further details on this model can be found in [4].



(a) Marginal posterior estimate for β_{62} in the 3 factor model

(b) Estimated model probabilities - dotted line shows gold standard

Figure: Performance with and without derivatives for posterior estimation and model choice.

Results: 100 SMC runs with $N = 10,000$ particles are performed. *MALA improves exploration for some of the more complex marginals (e.g. Figure (a) above), while using roughly half the number of likelihood evaluations of RW. The most accurate and precise model choice probabilities are obtained with MALA and ZV-CV. Again, we found that the SS and TI log evidence estimators are remarkably similar.*

Ongoing and Future Work

- MMALA - making use of the second derivatives
- control functionals for improved convergence
- an example with a nonlinear ordinary differential equation

References

1 Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539-552.

2 Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709-723.

3 Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123-214.

4 Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 41(1):41-67.

5 Marzolin, G. (1988). Polygynie du cincle plongeur (cinclus cinclus) dans le côtes de Loraine. *Oiseau et la Revue Francaise d'Ornithologie*, 58(4):277-286.

6 Mira, A., Solgi, R., and Imparato, D. (2013). Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653-662.

7 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 12(6):1087-1092.

8 Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numerical Mathematics*, 55(2):137-157.

This work was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers.